Predicting Biased Human Decision-Making with Large Language Models in Conversational Settings

Stephen Pilli
stephen.pilli@ucdconnect.ie
University College Dublin
Dublin, Ireland

Vivek Nallur
vivek.nallur@ucd.ie
University College Dublin
Dublin, Ireland

Abstract

We examine whether large language models (LLMs) can predict biased decision-making in conversational settings, and whether their predictions capture not only human cognitive biases but also how those effects change under cognitive load. In a pre-registered study (N = 1,648), participants completed six classic decision-making tasks via a chatbot with dialogues of varying complexity. Participants exhibited two well-documented cognitive biases: the Framing Effect and the Status Quo Bias. Increased dialogue complexity resulted in participants reporting higher mental demand. This increase in cognitive load selectively, but significantly, increased the effect of the biases, demonstrating the load-bias interaction. We then evaluated whether LLMs (GPT-4, GPT-5, and open-source models) could predict individual decisions given demographic information and prior dialogue. While results were mixed across choice problems, LLM predictions that incorporated dialogue context were significantly more accurate in several key scenarios. Importantly, their predictions reproduced the same bias patterns and load-bias interactions observed in humans. Across all models tested, the GPT-4 family consistently aligned with human behavior, outperforming GPT-5 and open-source models in both predictive accuracy and fidelity to human-like bias patterns. These findings advance our understanding of LLMs as tools for simulating human decision-making and inform the design of conversational agents that adapt to user biases.

CCS Concepts

• **Human-centered computing** → **Natural language interfaces**; **Empirical studies in HCI**; • **Computing methodologies** → **Discourse, dialogue and pragmatics**; **Interactive simulation**; **Artificial intelligence**; **Cognitive science**.

Keywords

Conversational AI, Framing Effect, Status Quo Bias, LLM Simulation

1 Introduction

Digital interfaces influence almost every aspect of modern life. They mediate interactions with loved ones, transport, medical checkups, entertainment, and sometimes even food and intimate choices. The function and form of these interfaces influence not only our view of the world, but also *how* we influence the world. Our decisions on which action to take, which option to ignore, and what aspect of a problem to pay attention to are all affected by the contextual elements within which the decision scenario appears. In particular, conversational interfaces now act as decision mediators across multiple domains, shifting the structure, presentation, and interpretation of available alternatives.

In its simplest form, decision-making requires a choice problem and a set of alternatives to choose from. The structure and presentation of these alternatives can potentially shape decision-making by tapping into underlying cognitive biases. Cognitive biases are systematic deviations from rational judgment, arising from heuristics, prior experiences, emotions, or social factors [28]. Over 200 such biases have been systematically cataloged and experimentally validated through standardized cognitive tasks [43]. While these biases have been studied extensively in static, survey-based, or GUI-mediated settings, relatively little is known about their manifestation in interactive, language-based environments like task-oriented dialogues. Early work suggests that biases persist even in conversational settings [2, 42, 60], yet we lack a systematic understanding of how the dynamics of dialogue shape bias susceptibility.

In decision theory, the process of decision-making requires the presence of an actor, denoted as the decision-maker, and a contextual environment within which the decision transpires [48]. Although cognitive biases originate from internal heuristics, they can be influenced by external environmental factors, such as cognitive load [7]. In conversational settings, such as task-oriented conversational agents, decision-making occurs within the context of a dialogue. In this setting, cognitive load arising from prior conversational context can potentially influence users' biased decision-making. We refer to such contextual influence as dialogue complexity. This complexity may act as a proxy for cognitive load, shaping the likelihood of biased decision-making. Leveraging the conversational context can help in accurately predicting when users are likely to be susceptible to cognitive biases. This opens the door to adaptive interventions that promote more informed, deliberate, and rational decisions.

Simulation and prediction of human behavior has long been a goal of research in Human-Computer Interaction (HCI), cognitive science, and behavioral modeling [25]. Large language models (LLMs) have been fine-tuned to produce fluent, human-like dialogue and frequently achieve high performance on established benchmarks [62]. This advancement presents opportunities beyond conventional applications, such as LLMs or generative agents, by enabling the simulation of large-scale human behavior in both experimental and policy contexts [1, 4, 24, 40, 41]. Of particular interest is the question of whether LLMs can serve as predictive models of human judgment and decision-making. That is, not just in mimicking language patterns, but simulating how contextual and cognitive factors drive biased behavior.

Prior work has primarily examined whether LLMs themselves exhibit cognitive biases when prompted to make decisions [15, 25, 33]. While informative, this line of research focuses on the presence of biases within LLMs, rather than their ability to model or predict human-biased behavior. A notable exception is Ying et al. [63], who used LLMs to simulate human decision-making but found substantial misalignment between model predictions and human rationality; however, cognitive biases were not the main focus of their work. Park et al. [41] introduced a method that enables behaviorally grounded predictions by constructing generative agents with rich, interview-derived memory representations. These agents have demonstrated strong predictive accuracy across surveys, personality assessments, and experimental tasks by simulating individual-level responses. These advances raise a key question: can LLMs simulate decision-making behavior that is not only human-like but also bias-sensitive and context-aware? Building on this approach, we investigate whether LLMs can simulate individuals given the chat transcripts such that the predictions on biased decision-making align accurately with those of their real-world counterparts.

To achieve this, we begin by investigating whether **cognitive biases manifest in conversational settings**, as prior research has primarily focused on isolated or survey-based tasks. This leads to our first research question **(RQ1)**: *Do established cognitive biases (Framing and Status Quo effects) manifest in conversational decision-making settings?* While this investigates the presence of biases, it does not account for how they may be shaped by the conversational context. To understand the **role of conversational context**, we next examine whether features like prior dialogue complexity systematically interact with the bias. This motivates **(RQ2)**: *How does prior dialogue complexity interact with cognitive bias susceptibility?* Building on this, we ask whether LLMs can **predict human decisions** across such contexts using limited information, forming **(RQ3)**: *Can LLMs predict individual human decisions using limited prior dialogue and demographic information?* The accuracy of individual-level prediction does not reveal whether LLMs capture the presence of cognitive biases in the population. A model might consistently choose the biased alternative, inflating accuracy while misaligning with the true distribution of human responses. To investigate this, we evaluate whether LLMs can reproduce population-level bias effects and their interaction with dialogue complexity. Therefore, we finally explore whether LLMs can **simulate collective behavior** by reproducing not only the presence of biases but also how they interact with dialogue complexity. This brings us to **(RQ4)**: *Can LLMs reproduce both the presence of cognitive biases and their interactions with dialogue complexity at the population level (collective behavior)?*

This paper addresses these questions through two empirical studies. First, we conduct controlled human-subject experiments (N = 1,648) using six well-established choice problems adapted for conversational settings. The choice problems are chosen to investigate prominent Framing and Status Quo effect cognitive biases. These cognitive biases are chosen as they are well-studied and replicated in the HCI, psychology, and behavioral economics literature. These studies systematically manipulate choice problems and prior dialogue complexity to examine their influence on Framing and Status Quo biased decision-making. The methods and findings are detailed in the Human Experiments section (3). Building on these human results, we then evaluate multiple LLM families by prompting them to simulate human decision-making under identical conditions. Specifically, we assess LLMs' ability to predict both individual-level decisions and sample-level bias patterns using participant demographics and dialogue transcripts, along with ablation analyses. The methodology and results are presented in the LLM Experiments section (4). Finally, the Discussion section (5) outlines results and discusses the implications for bias-aware interaction design and LLM Simulation for HCI.

2 Related Work

We review the related works in the following two key areas: a) cognitive bias and load in conversational agents, and b) LLM behavioral modeling capabilities.

2.1 Cognitive Biases in Decision-making Facilitated through Conversational Agents

Cognitive biases in human decision-making have been examined closely in the fields of Psychology, Behavioral Economics, and Human-Computer Interaction. Recent research has examined their potential to both leverage and mitigate such biases [10] using conversational agents. Ji et al. [26] studied cognitive biases in spoken conversational search (SCS), highlighting biases like anchoring and confirmation bias in the absence of visual cues. Their framework is largely theoretical but sets the stage for future bias-mitigation strategies in voice-based systems. Pilli [42] used chatbots to assess cognitive biases like Framing effects and Loss aversion. Participants exhibited typical bias responses, confirming chatbots as valuable tools for bias detection and measurement. Yamamoto introduced "suggestive endings" in chatbot dialogue, based on the Ovsiankina effect [60]. This design prompted users to engage more deeply, ask follow-up questions, and reflect longer, enhancing cognitive engagement. Dubiel et al. [14] examined the role of synthetic voice fidelity in decision-making. They found that high-fidelity voices, through cues like pitch and pace, enhanced source credibility and triggered affect heuristics, subtly influencing user choices. Ali Mehenni et al. [2] explored children's susceptibility to tasks resembling cognitive tasks used to infer cognitive biases by conversational agents and robots using a modified Dictator Game. Their findings revealed a stronger influence from artificial interlocutors than humans, pointing to authority and social influence biases, especially among vulnerable users. Kalashnikova et al. [29] investigated linguistic nudges promoting ecological behavior. By leveraging biases such as Status Quo bias and social conformity, they showed that chatbots and robots were more persuasive than humans in shaping opinions.

2.2 Language and Cognitive Load in Dialogue Systems

Prior work has examined how cognitive load influences decision-making broadly (e.g., Deck and Jahedi [12], Sweller [50]). Khare et al. [30], who explored how internal characteristics of the choice problem, such as information overload and choice overload, can impact Status quo bias. However, their focus is on the structure of the alternatives themselves. Experimental studies have examined the

Framing effect under cognitive load and found supporting evidence for dual-process theory, suggesting that cognitive load increases the influence of framing on decision-making [7, 57] It has been demonstrated that cognitive load is inversely related to task performance facilitated by the chatbot. Schmidhuber et al. [47] explored how the use of a chatbot affects users' mental effort when interacting with a new software product, and also to what extent the use of a chatbot affects users' productivity. The results showed that chatbot users experienced less cognitive load. Similarly, Brachten et al. [8] showed that chatbots can reduce the cognitive load needed to complete various tasks. One effective strategy to minimize cognitive load is to avoid presenting long responses or requiring users to provide complex inputs. The cognitive load induced by the dialogue is dependent upon the dialogue design. By increasing the elements and interactions between the elements, mental demand increases in turn, leading to a higher cognitive load. A poorly designed dialogue can result in cognitive load. These studies have explored various aspects of conversational agents, like linguistic features, the affect caused by dialogue voice modulation, the length of dialogue, and their role in influencing decision-making. However, these studies focus on individual decision points or choice problems only; they do not account for how previous interactions, conversational context, or prior dialogue influence biases in subsequent decision-making.

2.3 LLM Behavioral Modeling

Current research establishes that LLMs can reproduce aggregate bias patterns [15, 25, 33] but has not systematically investigated whether these models can predict individual decision-making, based on conversational cues and demographic information, particularly under varying cognitive load conditions. While LLMs demonstrate human-like biases when prompted appropriately, their capacity for individualized behavioral prediction in conversational contexts remains largely unexplored. This gap motivates our investigation into whether LLMs can serve as predictive tools for biased human behavior in realistic conversational scenarios, moving beyond surface-level bias reproduction toward contextually sensitive individual behavioral modeling.

3 Human Experiments

In investigating the first two research questions, we design a dialogue that follows a real-world task-oriented dialogue structure and facilitates decision-making, yet ensures experimental control and ecological validity by standardizing dialogue content and controlling for confounding variables. A formal representation of our dialogue $\mathcal{D}$ with a choice problem and prior dialogue is as follows:

$$\mathcal{D} = \{u_1^{sys}, u_2^{usr}, \ldots \underbrace{u_{t-k}^{sys}, \ldots, u_{t-1}^{usr}}_{\text{Prior Dialogue}}, \underbrace{u_t^{sys}, u_{t+1}^{usr}}_{\substack{\text{Decision}\\ \text{Scenario}\\ \text{and}\\ \text{Response}}} \ldots\}$$

This section outlines the choice problems adapted from classical behavioral economics studies, explains their integration into the chatbot with consistency and experimental control, describes the preceding Simple and Complex Dialogues, and details the experiment design and procedure.

3.1 Choice Problems

The chatbot is designed to have introductory utterances like greetings $\{u_1^{sys}, .., u_3^{sys}, \ldots\} \in \mathcal{D}$, which are followed by prior dialogue $\{u_{t-k}^{sys}, \ldots, u_{t-1}^{usr}\} \in \mathcal{D}$, and which is then followed by a choice problem $u_t^{sys} \in \mathcal{D}$. A choice problem typically involves a decision-making problem accompanied by a set of alternatives from which participants must choose. In traditional experimental designs, a questionnaire format is used for a between-subjects experiment to investigate the biases. We adapt the same experiment design where a control group is presented with a version of the choice problem where alternatives are described neutrally, while the treatment groups encounter scenarios in which one option is explicitly framed. The effect of respective cognitive bias is determined by analyzing the statistical difference in participant responses $u_t^{usr} \in \mathcal{D}$ between these groups.

Our experiments adapted six choice problems, three targeting classic Framing effects (Risky-choice Framing [52], Attribute Framing [31], and Goal Framing [3]) and three Status quo bias scenarios (Budget allocation, Investment decisions, and College job offers) which are drawn from Samuelson and Zeckhauser [46]. These problems are well-established in the literature and have been reproduced in recent replication studies [7, 59].

3.1.1 *Framing Choice Problems.*

Our experiments used three choice problems, each representing a different type of framing effect: *Risky-choice*, *Attribute*, and *Goal framing*. The choice problem for Risky-choice framing effect was adapted from choice problems described in a replication study by Bogdanov et al. [7]. The original study was performed by Wang [56]. The choice problem follows the popular Asian Disease Problem by Tversky and Kahneman [53]. Participants choose between two plans (Plan A and Plan B). The participant's choices reveal the framing effect. The Attribute framing problem involved restaurant selection and was adapted from Kuang et al. [31], where the same distance was described either in miles (Space) or minutes (Time), showing that people's preferences change depending on how the information is framed. The Goal framing choice problem was adapted from Aravind et al. [3], which tested how different goal-based cues influence public transit adoption. We selected the normative frame, highlighting environmental sustainability to encourage eco-conscious decisions. Participants choose between two travel modes (Public Transit and Personal Car) for a 10-mile trip, with or without any sustainability-related information as a framing cue. We refer to the framing effect-related choice problems for the control group as "Framed", and for the experimental group as "Alternatively Framed." The complete set of framing choice problems is presented in Table 9 of Appendix A.

3.1.2 *Status Quo Choice Problems.*

The experiment used three decision-making scenarios or choice problems adapted from: Budget allocation (BA), Investment decision making (IDM), and College job offers (CJ) Samuelson and Zeckhauser [46]. These choice problems were selected due to their well-documented effects, serving as strong baselines for evaluation. These were additionally reproduced in the replication study by Xiao et al. [59]. Moreover, they represent domains that are both widely studied in behavioral economics and highly relevant to practical applications in chatbot-based e-commerce and decision support systems. Each choice problem was implemented in three conditions: a *neutral condition*, where the alternatives were presented equally with no status quo option, and two *Status Quo conditions*, where one of the alternatives was framed as status quo. The decision maker can move away from the status quo or stick to the status quo, which reveals the bias. While adapting to conversational style, we have made minor modifications to the original choice problems. We refer to the Status Quo-related choice problem condition for the control as "Neutral" and experimental conditions as "Status Quo A" or "Status Quo B" based on the alternative in the status quo position. The list of choice problems for all conditions is detailed in Appendix A.2. We made minor modifications to the choice problems, which we report in Appendix A.3.

3.2 Prior Dialogue

A key feature of decision-making in a conversational setting is the presence of dialogue that precedes the decision scenario or a choice problem. We refer to this as the prior dialogue, denoted as $\{u^{sys}_{t-k}, \ldots, u^{usr}_{t-1}\} \in \mathcal{D}$, where $\mathcal{D}$ represents the full dialogue. To investigate our second research question **(RQ2)** that is the complexity of prior dialogue can potentially play a role in shaping subsequent decision-making we designed two types of preference elicitation tasks: one that facilitates low-effort interaction, referred to as the *Simple Dialogue*, and another that is cognitively demanding, referred to as the *Complex Dialogue*. The design characteristics of these tasks are as follows:

3.2.1 Simple Dialogue. A preference elicitation task was used, where participants were engaged in a set of short binary (Yes/No) questions about preferences within a familiar domain. This dialogue design was inspired by the Schema-guided Dialogue (SGD) dataset introduced by Rastogi et al. [45] and aimed to simulate a natural, ecologically valid, low-effort interaction with the chatbot. Importantly, the Simple Dialogue used a conservative dialogue strategy: questions were direct, unambiguous, and did not require reasoning or memory beyond the current turn. This ensured that the mental effort required by the dialogue was minimal. Participants were instructed to answer each question directly and were prompted to enter "I don't know" for *prior dialogue attention check* on one attribute. A full list of domains and associated questions can be found in the Section B.1 of Appendix B. An example of the Simple Dialogue in the "Music" domain is shown below.

Table 1: Simple Dialogue Attributes and Respective Utterances

Attribute	Yes/No Question
Genre Preference	Do you like listening to pop music?
Language of Lyrics	Do you prefer music with lyrics in English?
Live Performances	Are you interested in live music performances?
Instruments Focused	Do you enjoy instrumental music?
Artist-Specific	Do you like music from specific artists? Please enter "I don't know" only.
Era (e.g., 80s, 90s)	Do you prefer music from the 90s?

3.2.2 Complex Dialogue. The Complex Dialogue was designed to induce cognitive load in a controlled yet ecologically valid manner. To achieve this, a nested referential structure is used to manage multiple interdependent entities across a multi-turn dialogue. The following shows an example of the utterances from u^{sys}_{t-k} to u^{sys}_{t-2} by the chatbot during the prior dialogue.

u^{sys}_{t-8}: The first artist performs three live shows, is paid 2000 units per show, and has a 4-star rating. The second artist performs twice as many shows, with the same pay and rating. Which artist do you prefer, and why?

u^{sys}_{t-6}: The third artist performs the same number of shows as the second, earns half the pay of the first artist, but has the same rating as the first. Which artist do you prefer, and why?

u^{sys}_{t-4}: The fourth artist performs the same number of shows as the second, earns the same pay as the third, but has two stars less than the first artist. Which artist do you prefer, and why?

u^{sys}_{t-2}: Remember the details of the fourth artist. Specific information will be requested later.

The fourth artist in the chat transcript is defined by the second and third artists, which reference the second and first artists, creating a chain of dependencies. This design necessitates users to maintain and integrate hierarchical relationships between entities, thereby increasing semantic integration costs and working memory demands. This was designed based on the psycholinguistic findings that nested dependencies elevate processing difficulty [19, 55]. Additionally, the reappearance of earlier referents after intervening turns results in high referential distance [11], which further taxes memory retrieval processes [5]. From a cognitive load design standpoint, this structure directly aligns with Sweller's Cognitive Load Theory (CLT), which distinguishes between intrinsic load (task-related complexity), extraneous load (inefficient information presentation), and germane load (effort used for schema building) [12, 50].

Our design increases intrinsic load by requiring participants to track and integrate multiple interrelated referents, and germane load by promoting mental model construction to resolve semantic dependencies across turns. From a dialogue systems perspective, managing multiple entities simultaneously while maintaining coherent context is a well-documented challenge, particularly when entities are interconnected or revisited [17, 54]. This reflects real-world conversational demands where dialogue agents must track, differentiate, and link multiple referents simultaneously. Our Complex Dialogue task is a preference elicitation task that necessitates arithmetic comparisons between attributes and memorization of outcomes, requiring additional mental effort. Similarly designed dialogues are used in other domains, including artist recommendations, streaming services, calendar apps, and banking options. (For the remaining complex dialogues, please refer to Table 11 in Section B.2 of Appendix B).

The domains of these tasks are adopted from the Schema-Guided Dialogue (SGD) dataset [45]. In theory, the task design must substantially increase the cognitive load of the individual. To empirically verify the cognitive load, the standard NASA-TLX [39] survey was adopted for all the interactions. To evaluate the cognitive load of the prior dialogue, we incorporated two indicators: the self-reporting

NASA-TLX and a *Recall Task*. The NASA-TLX was recorded after the chatbot interactions to capture participants' perceived cognitive load across multiple dimensions, such as mental demand and effort. Simultaneously, the Recall Task served as a behavioral indicator of attention and memory. Participants were asked to recall memorized arithmetic from the conversation. Additionally, we also investigate the *familiarity* of the participants with prior dialogue domains as a confounding factor.

3.3 Chatbot Technical Details

The web-based experimental interface was developed using Streamlit [49], which allowed for easy deployment and consistent access across devices. The source code for the chatbots used in the experiments is publicly available. The chatbot used in the framing experiment can be found at https://github.com/stephen-pilli/PEM.git, while the chatbot used in the status quo bias (SQB) experiment is available at https://github.com/stephen-pilli/exp-status-quo-bias.git.

GPT-4o Mini, a large language model [37], was used to create realistic and coherent chatbot interactions based on structured prompts designed for each condition. An example of the prompt used for the agent (LLM chatbot) is provided in Appendix H.

3.4 Design of the Experiments

We employed a two-factor between-subjects design. The first factor is the prior dialogue complexity, which manipulated the cognitive load experienced by participants before making a decision. The second factor is the choice problem condition (Recall the conditions for Framing and Status quo detailed in Section 3.1).

For the Framing experiments, we employed a 2 × 2 experiment design. The first factor was dialogue complexity with two levels (Simple vs. Complex), and the second factor was framing with two levels (Framed vs. Alternatively Framed). Participants were randomly assigned to one of the four conditions, ensuring balanced group sizes. For the Status quo experiments, we employed a 2 × 3 factorial design. Dialogue complexity again had two levels (Simple vs. Complex), while the Status Quo factor had three levels (Neutral, Status Quo A, Status Quo B). Participants were randomly assigned to one of the six resulting conditions, with balanced distribution across groups. To preserve internal validity and avoid carryover effects, each participant encountered only one version of the dialogue and one framing condition.

The primary dependent variable in our study is participants' choices between alternatives in the choice problems. All six choice problems were included in the experiment. Decision outcomes were compared across bias conditions to address our research questions. Our second key variable is a *moderator*: the cognitive load required to complete the task. We measured cognitive load using the NASA-TLX questionnaire as well as behavioral indicators. This variable allowed us to statistically test whether complex dialogues place greater mental demands on participants compared to simple dialogues. To avoid confounding effects, the domains of the prior dialogues and the choice problems were intentionally different. This separation ensured that cognitive load was isolated from other factors, such as domain familiarity, and allowed us to focus on how dialogue complexity influenced subsequent decision-making.

While our dialogue tasks were adapted from classic behavioral economics experiments, we carefully designed them to preserve ecological validity by embedding the decision-making within naturalistic, task-oriented chatbot interactions. This ensured that participants experienced the scenarios as they would in a real conversational setting with an intelligent agent, rather than as isolated survey questions. At the same time, we maintained experimental control by standardizing dialogue length, turn-taking, and framing conditions across participants. This allowed us to capture bias-prone decision-making in a realistic human–agent interaction context, while still ensuring internal validity and replicability of the results.

3.5 Power Analysis, Recruitment, and Data Integrity

We conducted an *a priori* power analysis using G*Power [18], targeting 0.80 power to detect a medium effect size (ω = 0.3, α = 0.05) following Pancholi et al. [38]. This required approximately 42 participants per condition (see Appendix C). Participants (N=1648) were recruited via Prolific [44], compensated at $8/hour, and randomly assigned to 2x2 (Framing) or 2×3 (Status Quo) designs. The studies were preregistered on the Open Science Framework (OSF). The preregistration for Framing study is archived at https://doi.org/10.17605/OSF.IO/DPR45, and the preregistration Status quo study is archived at https://doi.org/10.17605/OSF.IO/PSXVF. Data integrity was ensured through attention checks, recall tasks, and automated JSON-based logging. The dataset for Framing effect study is available at https://doi.org/10.5281/zenodo.18218753, and the Status quo bias study is available at https://doi.org/10.5281/zenodo.16541481.

3.6 Procedure

Participants began the study by reviewing an information sheet outlining the study's purpose, procedures, and ethical considerations. After reading the document, they were directed to the experiment's homepage, where their Prolific ID (Stored in an irreversible, anonymous state) was displayed alongside the consent form.

To ensure active participation, the consent checkbox was initially unchecked, requiring participants to explicitly select "Yes" before proceeding. Only after providing informed consent were they granted access to the experiment. Participants interacted with the chatbot based on the condition they are randomly assigned to as shown in the Figure 1. After completing the dialogue with the chatbot, participants were redirected to a questionnaire containing additional measures to ensure data quality. The survey included a memory recall task to assess attentiveness, along with attention check questions to verify engagement. By implementing these steps, the study ensured that participants remained actively engaged and provided high-quality responses, ultimately enhancing the reliability of the collected data. After each task, participants reviewed their transcript and then completed the NASA-TLX survey. In both the Simple and Complex Dialogue conditions, participants rated their perceived mental demand. This design minimized recall bias and enabled a precise assessment of the mental demands imposed by dialogue complexity.

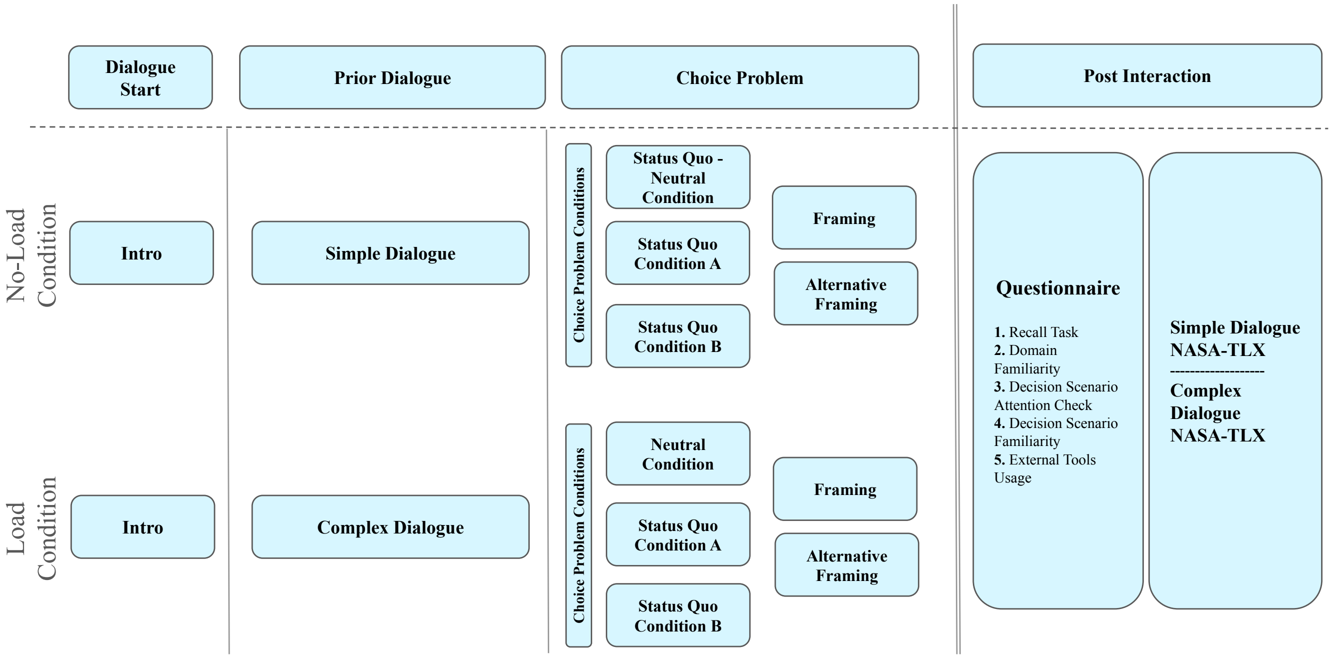

Figure 1: Experimental procedure outlining the sequence of tasks. Participants were first introduced to the dialogue (Simple or Complex), followed by the assigned condition (Framing: Framed vs. Alternative Framing; Status Quo: Neutral, A, or B). After completing the choice problem, participants filled out the NASA-TLX and post-task questionnaires (recall, familiarity, attention checks, and tool usage).

3.7 Findings

3.7.1 Dataset Description. For the Framing study, the initial sample of 595 participants was filtered based on familiarity ratings, attention checks, prior exposure to the choice problem, invalid responses, and missing NASA-TLX data, resulting in a final sample of 548. The mean age was 43 years (SD = 13.7). The dataset is a slightly female-skewed sample (53.5%) and a majority identifying as White (86.5%). Most participants were not students (72.1%) and were employed either full-time (46.5%) or part-time (18.1%). Detailed demographic distributions across experimental conditions are presented in Table 12 in Appendix D.

For the Status quo study, an initial sample of 1,256 participants underwent data cleaning, which excluded 19 for invalid responses, 40 for failing scenario recall, 49 for prior familiarity, and 63 for using external tools, resulting in a final sample of 1,100 participants. The mean age was 41.5 years (SD = 13.3), with an even gender distribution (50.5% female). Most participants were residents of the United Kingdom (n = 778), followed by the United States (n = 308) and Ireland (n = 14). Full demographic breakdown for the Status quo experiment is provided in Table 13 in Appendix D.

3.7.2 Framing and Status Quo effects reproduced in Conversational Setting. Table 2 presents evidence across six choice problems that addresses ***RQ1*** and ***RQ2***. ***RQ1*** investigates whether Framing and Status quo effects can be reproduced in a conversational setting; several strong and moderate effects were observed. Risky-choice framing showed a strong effect under complex dialogue but only a weak effect under simple dialogue, which is comparable to the original study by Wang [56]. Attribute framing showed a robust effect in the literature [31]; however, the same did not replicate strongly in this conversational setting; only weak evidence appeared under simple dialogue, and no effect under complex dialogue. Goal framing produced the clearest result, with a strong effect both in the original study and under complex dialogue, and a weak effect in simple dialogue. The Status quo scenarios (budget allocation, investment decisions, and college jobs) showed mixed evidence: budget allocation and college jobs replicated strongly under both dialogue types, while the investment decision scenario showed no effect. Overall, these results address ***RQ1***, demonstrating that several well-documented Framing and Status Quo effects persist in conversational settings, though their strength varies across the choice problems.

3.7.3 Prior Dialogue Complexity Resulted in Cognitive Load. Across both the Framing and Status quo experiments, NASA-TLX results consistently showed that Complex dialogues resulted in significantly higher cognitive load than Simple dialogues. Mental Demand increased most strongly ($d = 0.85 - 1.08, p < .001$), followed

Table 2: Summary of human experiment results. The table reports effect sizes (Cohen's h with 95% CI), p-values, and significance tests for detecting cognitive biases (H1) and their interaction with dialogue complexity (H2). Results are organized by cognitive bias (Framing Effect, Status Quo Bias) and choice problem. The tickmark (✓) indicates evidence supporting the hypothesis, while cross (✗) indicates no support.

Cognitive Bias	Choice Problem	Study	Cohen's h [95% CI]	p-value	Bias Found (RQ1)	Interaction With Dialogue Complexity (RQ2)
Framing Effect	Risky Choice	Wang [56]	0.193 [-0.061, 0.447]	0.458	✗	Positive
		Simple Dialogue	0.205 [-0.001, 0.410]	0.392	✗	
		Complex Dialogue	0.730 [0.524, 0.937]	0.001	✓	
	Attribute	Kuang et al. [31]	0.267 [0.2, 0.335]	< 0.001	✓	Negative
		Simple Dialogue	0.291 [0.093, 0.489]	0.158	✗	
		Complex Dialogue	0.135 [-0.060, 0.331]	0.549	✗	
	Goal	Aravind et al. [3]	0.675 [0.606, 0.744]	< 0.001	✓	Positive
		Simple Dialogue	0.225 [0.019, 0.432]	0.388	✗	
		Complex Dialogue	0.567 [0.360, 0.774]	0.014	✓	
Status Quo Bias	Budget Allocation	Samuelson and Zeckhauser [46]	0.78 [-0.26, 1.82]	0.025	✓	No Interaction
		Simple Dialogue	0.779 [0.602, 0.956]	< 0.001	✓	
		Complex Dialogue	0.794 [0.618, 0.969]	< 0.001	✓	
	Investment	Samuelson and Zeckhauser [46]	0.38 [-0.21, 0.98]	0.069	✗	No Interaction
		Simple Dialogue	0.082 [-0.101, 0.266]	0.666	✗	
		Complex Dialogue	0.043 [-0.145, 0.231]	1	✗	
	College Jobs	Samuelson and Zeckhauser [46]	1.26 [0.31, 2.21]	< 0.001	✓	No Interaction
		Simple Dialogue	0.463 [0.284, 0.642]	0.014	✓	
		Complex Dialogue	0.577 [0.396, 0.758]	0.002	✓	

by Effort ($d = 0.6 - 0.77, p < .001$), with smaller but reliable increases for Performance, Frustration, and Temporal Demand, while Physical Demand showed minimal effects. Behavioral indicators supported these findings: participants in the Complex condition took longer to respond and demonstrated higher accuracy on the memory recall task, both correlating positively with self-reported Mental Demand. Together, these converging results confirm that complex prior dialogue substantially increased cognitive load, validating our manipulation across both experimental studies. Behavioral indicators further validated our cognitive load manipulation. In the framing study, accuracy correlated positively with Mental Demand ($r = 0.13, p = 0.002$), showing that participants who remembered task details also reported higher workload. In the Status Quo study, recall accuracy correlated with both response time ($r = 0.318, p < .001$) and Mental Demand ($r = 0.105, p = .013$), while participants in the Complex Dialogue took significantly longer than in the Simple Dialogue ($d = 0.59, p < .001$). Together, these results demonstrate that complex prior dialogues consistently increased cognitive load. For detailed exposition, please refer to the Appendix Section E.

3.7.4 Interaction between Complex dialogue and biased decision-making. Our second research question investigated whether prior dialogue complexity interacts with subsequent decision-making; the results suggest selective but meaningful interactions. In Risky-choice framing and Goal framing, effect sizes were significantly larger following complex dialogue than simple dialogue, with confidence intervals indicating strong interactions. This implies that complex prior dialogue increases the susceptibility to these effects, a pattern consistent with findings from randomized control trials in psychology [7, 57]. In contrast, Attribute framing showed a negative interaction, but it is not significant. Budget allocation, investment, and college jobs showed no interaction between dialogue complexity and decision outcomes. The effects either remained stable across the choice problem and prior dialogue. Taken together, these findings suggest that prior dialogue complexity can strengthen the Framing effect but does not affect Status Quo bias, addressing our Research Question (***RQ2***).

4 LLM Experiments

To evaluate whether large language models (LLMs) can reproduce observed biased human decision-making patterns by predicting at the individual level, we replicated the experimental method used with human participants with LLMs. Each LLM was prompted using two key inputs: (a) the demographic attributes available from Prolific (e.g., age, gender, education, and country of residence) and

(b) the transcript of the participant's dialogue up to the choice problem or choice problem; this chat includes the prior dialogue, as shown in the Chat Transcript 4.1. LLMs presented with the same choice problems as their corresponding human participants and were asked to act as participants and predict the decision the participant would make. This design ensured that the information available to the LLMs mirrored the information grounding human decisions.

4.1 Human-Likeness Prompts

Chat Transcript

User: You are a participant in a research study. The following is your demographic information: {‘Age’: 25, ‘Sex’: Female, ‘Ethnicity simplified’: Black, ‘Country of birth’: United Kingdom, ‘Country of residence’: United Kingdom, ‘Nationality’: United Kingdom, ‘Language’: English, ‘Student status’: None, ‘Employment status’: Full-Time}. Your task is to respond to the user. Study Instructions: The user will ask various questions to understand your preferences on a topic. At some point, the user will change the topic and ask you a different scenario-based question. Please pay attention and ensure you understand the question before answering. Later in this study, you must answer three extra questions to identify if you understood the scenario-based question and all the options correctly. After the scenario-based question, the user will present various attributes. You need to compare and select the best. When the user asks you to remember specific details, please memorise them carefully, as you will need to recall them later in the study. Shall we begin? Please respond with ’yes’ to start the conversation.
Assistant: Yes, let's begin!
User: Hello! I'm here to understand your preferences through various Scenarios. Shall we start?
Assistant: Yes, lets begin
User: Do you like listening to pop music?
Assistant: Yes i do
⋮
User: Do you prefer music from the 90s?
Assistant: A little
User: Thanks for sharing! Now, let's get started with some specific scenarios to understand your preferences. The first artist performs three live shows, is paid 2000 units per show, and has a 4-star rating. The second artist performs twice as many shows, with the same pay and rating. Which artist is better, and why?
Assistant: The second artist
⋮
User: Remember number of live performances, remuneration, and the star rating of the fourth artist. Now, let's move on to a different scenario. You are currently an assistant professor at College A in the east coast. Recently, you have been approached by a colleague at another university with a job opportunity. When evaluating teaching job offers, people typically consider the salary, the reputation of the school, the location of the school, and the likelihood of getting tenure (tenure is a permanent job contract that can only be terminated for cause or under extraordinary circumstances). Your choices are:

- Remain at College A: east coast, very prestigious school, high salary, fair chance of tenure.
- Move to College B: west coast, low prestige school, high salary, good chance of tenure.

Assistant: ⋯

A central design choice in the LLM experiments was how to instruct LLMs to simulate human participants. Following recent work on LLM behavioral prompting [6, 32, 63], we adopted a series of human-likeness prompts that varied in the degree to which the LLM was instructed to emulate human reasoning. Prior research suggests that LLMs may default to more rational or normatively consistent behavior than humans [32]. The extent of instruction required for an LLM to align with actual human behavioral patterns is a latent variable that requires systematic investigation. To explore this, we varied the level of human-likeness in the prompts, ranging from minimal role prompts to explicit directives to exhibit susceptibility to cognitive biases.

Specifically, we implemented three levels of human-likeness. At human-likeness Level 1 (HL1), LLM received only a minimal role instruction: “You are a participant in a research study.” This established a research setting without explicit guidance on how to respond, allowing us to assess the LLM's baseline behavior. At human-likeness Level 2 (HL2), LLMs were encouraged to simulate more naturalistic human responses with the prompt: “You are a human participant in a research study. Please answer questions as naturally as you would in everyday life.” This formulation aimed to elicit more ecologically valid answers while avoiding explicit mention of biases. At human-likeness Level 3 (HL3), we explicitly instructed the LLM to act as humans prone to cognitive biases: “You are a human participant in a research study. Therefore, act as a human. Be highly susceptible to cognitive biases such as Framing, Status Quo bias, and Anchoring when reasoning and answering questions. Avoid overthinking and lean into intuitive, sometimes irrational judgments.” This highest level of human-likeness was designed to test whether LLMs could be guided to reproduce not just biases but also the susceptibility to cognitive load.

4.2 Technical Details

All LLM simulations were conducted using a combination of proprietary and open-source models. Proprietary models (GPT-4.1, GPT-4.1-mini, GPT-5, and GPT-5-mini) were accessed via the OpenAI API using the chat.completions.endpoint. Open-source models (gpt-oss-120b, llama4, and qwen3) were run on Google Cloud. To ensure reproducibility, we employed batch mode execution, fixed the random seed to 42, and set the temperature parameter to 0 to enforce deterministic outputs. For OpenAI models, we additionally logged the system fingerprint returned by the API to track model versions. Annotation of the output was conducted separately using three different LLMs: GPT-4.1, GPT-4.1-mini, and GPT-5-mini. Interrater Agreement (IRA) was calculated to choose the annotation for analysis.

4.3 Findings

4.3.1 LLM predictions of biased human decision-making. To address ***RQ3***, can LLMs predict individual human decisions using limited prior dialogue and demographic information? We evaluated how accurately LLMs predicted participants' choices and how prior dialogue contributed to these predictions. To understand the contribution of prior dialogue to LLM prediction performance, we compared accuracy across three conditions: *Choice Problem Only*, *Without Prior Dialogue*, and *With Prior Dialogue*. In the *Choice Problem Only* condition, the LLM was provided with just the text of the decision-making scenario (e.g., a Framing or Status Quo choice task), without any additional context such as demographics or prior dialogue. This serves as a baseline condition and assesses whether the model can predict the participant's choice based solely on the choice problem itself. In the *Without Prior Dialogue* condition, the model was provided with demographic information (e.g., age, gender, education, and country of residence) along with a human-likeness prompt instructing the model to respond in a human-like manner. However, the prior dialogue was withheld. This allows us to isolate the effect of demographics and role-prompting on prediction performance. Finally, in the *With Prior Dialogue* condition, the model received all available contextual input, including demographics, human-likeness prompt, and the full transcript of the dialogue prior to the choice problem. This condition enables us to evaluate whether the LLM uses prior dialogue for prediction.

LLM prediction accuracy varied substantially across choice problems and dialogue conditions. Three distinct patterns emerged:

(1) ***Case 1.*** No Dialogue Effect: For some problems (e.g., Risky Choice and Attribute Framing), including prior dialogue or demographic prompts did not significantly change prediction accuracy, suggesting that dialogue context added little predictive value.
(2) ***Case 2.*** Dialogue-Enhanced Prediction: In other problems (notably Goal Framing and Investment Decisions), accuracy improved markedly when prior dialogue was included. For instance, in the Goal Framing, accuracy increased from 47% (*Without Prior Dialogue*) to 63% (*With Prior Dialogue*) decisions. Similarly, for Investment decision-making, accuracy increased from 62% to 76% as shown in Table 3. This shows that conversational context can align LLM predictions more closely with actual human decision-making.
(3) ***Case 3.*** Consistently High Accuracy: Some tasks (e.g., Budget Allocation and College Jobs, both involving Status Quo bias) achieved high accuracy across all conditions, indicating stable human preferences that LLMs could capture even without dialogue context.

These results address ***RQ3*** by demonstrating that LLMs can predict individual human decisions more accurately when provided with conversational context in addition to demographic information. However, the extent of this improvement varies depending on the type of choice problem. The complete set of results and detailed analysis is provided in the Table 14 of Appendix F.

4.3.2 Biases observed in human experiments reproduced at the sample level. Table 4 presents a comparison of human and LLM behavior across six choice problems under varying levels of human-likeness. Various choice problems demonstrated clear evidence of bias in both Simple and Complex dialogue for human experiments. Investment Decision Making (IDM) showed no effect (original study had marginal significance [46], while Attribute Framing (ATF) showed no significant effect in either condition. Findings in the original study of ATF often yield weak or inconsistent biases showing small effect sizes.)

LLMs displayed varying levels of bias across the three human-likeness conditions. When explicitly instructed to behave in a biased manner (Human-Likeness 3: “You are a human participant in a research study. Therefore, act as a human. Be highly susceptible to cognitive biases such as framing, status quo bias, anchoring”), LLMs showed strong bias across all choice problems, including Attribute Framing (ATF) and Investment Decision Making (IDM), where no bias was observed in the human experiments. This resulted in false positives, particularly in HL3 (shown in Table 5), revealing a forced, biased behavior when asked. Consequently, the accuracy (distinct from individual-level prediction accuracy used in Section 4.3.1 to report individual-level prediction), which is the proportion of choice problems in LLM experiments that matched actual human experiments, was only 58% in HL3. In contrast, when LLMs were given a more neutral prompt (“You are a participant in a research study”), accuracy improved to 75%. Similar accuracy was achieved in Human-Likeness 2, where agents were instructed: “You are a human participant in a research study. Please answer questions as naturally as you would in everyday life.” Unlike HL3, this prompt avoided explicitly referencing cognitive biases. Under HL1 & HL2, GPT4.1 correctly reproduced the absence of bias in both attribute Framing and the Investment Status Quo choice problem, closely aligning with human behavior and reducing false positives.

4.3.3 LLMs reproduce observed biased human behavior under complex prior dialogue. We conducted independent t-tests comparing effect sizes between Simple and Complex dialogue conditions. In the human data, Risky Choice Framing (RCF) and Goal Framing (GF) bias effects were significantly stronger after complex dialogues, consistent with prior research in cognitive psychology that links increased mental load with greater reliance on intuitive or biased decision-making. However, in the Status Quo bias scenarios (e.g., Budget Allocation and College Jobs), although bias was present in both conditions, the effect sizes remained relatively stable between Simple and Complex dialogues, indicating little or no interaction with cognitive load.

To investigate whether LLMs could capture this interaction pattern (***RQ4***), we examined the direction and magnitude of effect size changes across dialogue conditions using z-scores. For example, in GF, the effect size for humans increased from 0.225 (Simple) to 0.567 (Complex) (as shown in Table 4), resulting in a positive z-score of 2.29 (as shown in Table 6), indicating a stronger bias under cognitive load. However, in HL1, the LLM's effect size decreased from 2.29 to 1.976 between conditions, resulting in a negative z-score of -1.61, meaning that the LLM behaved in the opposite direction. Interestingly, HL3 showed a positive z-score of 2.66, indicating that under cognitive load, LLM's responses were more biased, similar to human responses. Similar trends were observed for complementary cases like Attribute Framing (ATF) and Investment, where humans showed a negative direction (less bias under complexity), which was only mirrored correctly by HL3 but not HL1 or HL2.

To investigate the direction and magnitude statistically, we calculated Spearman correlation ρ between the human experiment z-scores and those of each LLM human-likeness condition. The results revealed a marginally significant positive correlation between Human and HL3 (ρ = 0.771, p = .07). In contrast, HL1 and HL2 showed weak correlations (ρ = 0.600), indicating a poor match with human behavior in terms of representing how dialogue complexity interacts with decision-making.

Overall, our findings show that LLMs can reproduce sample-level human biases such as Framing and Status Quo Bias, especially under neutral prompting conditions (HL1 and HL2), achieving up to 75% alignment with human responses. However, under HL3, where models were explicitly told to simulate bias, they overestimated effects, leading to false positives. LLMs struggled to reproduce load-bias interactions, such as the impact of cognitive load, unless explicitly prompted, like in HL3.

Table 3: GPT-4.1 prediction accuracy across decision problems and dialogue conditions. Values show mean accuracy (two decimals). Asterisks denote significant improvement over the Choice Problem Only condition (* *p* < .05, ** *p* < .01, * *p* < .001). Cases correspond to patterns observed: Case 1 = No Dialogue Effect, Case 2 = Dialogue-Enhanced Prediction, Case 3 = Consistently High Accuracy.**

Bias Type	Choice Problem	Choice Only	No Dialogue	With Dialogue	n	Case	Comparing No Dialogue and With Dialogue
Framing Effect	Risky Choice	.52	.62	.60	177	1	Stable pattern
	Attribute	.47	.48	.48	195	1	Stable pattern
	Goal	.34	.47	**.63*****	176	2	Significant improvement
Status Quo Bias	Budget Allocation	.72	.72	.72	377	3	Consistently high accuracy
	Investment	.28	.62	**.76*****	327	2	Significant improvement
	College Jobs	.61	.56	.53	197	3	Stable pattern

Table 4: Human and model effect sizes (Cohen's h) with 95% confidence intervals. Significance: * p < 0.05, ** p < 0.01, *** p < 0.001.

Choice Problem	Prior Dialogue	Human	LLM		
			Human Likeness 1	Human Likeness 2	Human Likeness 3
Risky Choice Framing	Simple Dialogue	0.205 [-0.001, 0.410]	1.770 [1.565, 1.976]***	1.252 [1.047, 1.458]***	0.690 [0.485, 0.896]**
	Complex Dialogue	0.730 [0.524, 0.937]**	2.447 [2.238, 2.656]***	2.299 [2.092, 2.505]***	0.810 [0.602, 1.017]***
Attribute Framing	Simple Dialogue	0.291 [0.093, 0.489]	0.648 [0.450, 0.846]**	0.902 [0.704, 1.100]***	1.170 [0.972, 1.368]***
	Complex Dialogue	0.135 [-0.060, 0.331]	0.244 [0.048, 0.439]	0.115 [-0.080, 0.311]	0.928 [0.732, 1.123]***
Goal Framing	Simple Dialogue	0.225 [0.019, 0.432]	2.216 [2.009, 2.423]***	2.089 [1.883, 2.296]***	0.347 [0.140, 0.554]
	Complex Dialogue	0.567 [0.360, 0.774]*	1.976 [1.769, 2.182]***	1.820 [1.613, 2.026]***	0.744 [0.537, 0.950]**
Budget Allocation	Simple Dialogue	0.779 [0.602, 0.956]***	1.503 [1.325, 1.680]***	0.990 [0.813, 1.168]***	2.781 [2.603, 2.958]***
	Complex Dialogue	0.794 [0.618, 0.969]***	0.891 [0.715, 1.066]***	0.596 [0.421, 0.772]**	2.783 [2.608, 2.959]***
Investment	Simple Dialogue	0.082 [-0.101, 0.266]	0.147 [-0.036, 0.331]	0.007 [-0.177, 0.190]	2.766 [2.583, 2.950]***
	Complex Dialogue	0.043 [-0.145, 0.231]	0.009 [-0.179, 0.197]	0.009 [-0.179, 0.197]	2.758 [2.570, 2.946]***
College Jobs	Simple Dialogue	0.463 [0.284, 0.642]*	2.776 [2.597, 2.955]***	2.777 [2.598, 2.956]***	2.193 [2.014, 2.373]***
	Complex Dialogue	0.577 [0.396, 0.758]**	2.773 [2.592, 2.954]***	2.773 [2.592, 2.954]***	2.625 [2.444, 2.806]***

Table 5: Confusion matrices for HL1, HL2, HL3 (Biased/Not Biased). We capture true positives, true negative, and false positives therefore accuracy as a metric explains our findings better.

	LLM (HL1)	
Human	Not Biased	Biased
Not Biased	3	3
Biased	0	6
	Accuracy	0.75

	LLM (HL2)	
Human	Not Biased	Biased
Not Biased	3	3
Biased	0	6
	Accuracy	0.75

	LLM (HL3)	
Human	Not Biased	Biased
Not Biased	1	5
Biased	0	6
	Accuracy	0.58

4.4 Analysis Across Models

Sample-level accuracy shows how often the LLMs correctly reproduce human biases. A higher accuracy means the LLM accurately reproduced biased behavior in our human experiment, while lower accuracy indicates the LLM tends to exhibit bias where humans do not exhibit, or vice versa. Z-score captures the change in the effect size of a bias under complex prior dialogue. Spearman correlation uses the Z-scores to test the monotonic relation between the human experiments and the LLM. A strong positive correlation suggests the LLM reproduced the changes in effect size of biases, showing

Table 6: Comparison of z-values and confidence intervals for effect size differences (Simple vs Complex Dialogue). Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Choice Problem	Human	HL1	HL2	HL3
Risky Choice Framing	3.53 [0.23, 0.82]***	4.53 [0.38, 0.97]***	7.04 [0.76, 1.34]***	0.81 [-0.17, 0.41]
Attribute Framing	-1.10 [-0.43, 0.12]	-2.85 [-0.68, -0.13]**	-5.54 [-1.07, -0.51]***	-1.70 [-0.52, 0.04]
Goal Framing	2.29 [0.05, 0.63]*	-1.61 [-0.53, 0.05]	-1.81 [-0.56, 0.02]	2.66 [0.10, 0.69]**
Budget Allocation	0.12 [-0.23, 0.26]	-4.81 [-0.86, -0.36]***	-3.09 [-0.64, -0.14]**	0.02 [-0.25, 0.25]
Investment	-0.29 [-0.30, 0.22]	-1.03 [-0.40, 0.12]	0.01 [-0.26, 0.26]	-0.06 [-0.27, 0.25]
College Jobs	0.88 [-0.14, 0.37]	-0.02 [-0.26, 0.25]	-0.03 [-0.26, 0.25]	3.32 [0.18, 0.69]**

the sensitivity to cognitive load, whereas weak or negative correlations indicate a misalignment with human biased decision-making patterns under load, as observed in our experiments.

Table 7: Sample Level Accuracy and Spearman correlation (ρ) for each model. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{\dagger}p < 0.10$ (marginal significance).

Model	Accuracy			Correlation (ρ)		
	HL1	HL2	HL3	HL1	HL2	HL3
gpt4.1-mini	0.833	0.750	0.667	0.290	0.371	0.543
gpt4.1	0.750	0.750	0.583	0.600	0.600	0.771†
gpt5-mini	0.667	0.583	0.667	-0.429	-0.143	-0.314
gpt5	0.333	0.417	0.667	$-0.771^{\dagger}$	-0.714	-0.314
gpt-oss-120b	0.667	0.583	0.667	-0.143	-0.143	0.143
llama4	0.417	0.583	0.583	-0.029	-0.257	-0.714
qwen3	0.583	0.667	0.500	0.257	-0.371	0.257

Table 7 summarizes the accuracy and correlation scores for each model across the three human-likeness levels. Across all models, GPT-4.1 consistently showed the best performance. Its accuracy was moderately high for both H1 and H2 (0.75), and still reasonable for H3 (0.58). It also showed a marginally significant positive correlation with human data in H3 (ρ = 0.771, p < .10), suggesting it could reproduce both the presence of biases and their change under cognitive load. GPT-4.1-mini had slightly higher accuracy (0.833 for H1, 0.750 for H2, and 0.667 for H3), but it did not show meaningful correlations, limiting its interpretability. In contrast, the GPT-5 family performed poorly. GPT-5-mini had moderate accuracy (0.580–0.670 across H1–H3), but its correlations were near zero or negative, meaning it often missed the direction of change in bias. GPT-5 had the weakest results, with low accuracy (0.333–0.670) and a marginally significant negative correlation in H2 (ρ = −0.771, p < .10), indicating it often predicted the opposite of human behavior under cognitive load. Among open-source models, performance was mixed but generally weaker. GPT-OSS-120b showed moderate accuracy (0.583–0.667) but failed to capture bias–load interactions. LLaMA4 had lower accuracy (0.417–0.583) and consistently negative correlations (−0.029 to −0.714), suggesting strong divergence from human-like behavior. Qwen3 performed inconsistently, with decent accuracy in HL2 (0.667) but weak results elsewhere. Overall, GPT-4.1 was the most aligned with human behavior, especially under cognitive load. The GPT-5 family often misrepresented bias patterns, while open-source models showed limited ability to simulate human-like decision-making, particularly in dynamic or context-sensitive settings.

4.5 Ablation & Perturbation

To better understand which components of our experimental setup contributed to the LLM's ability to reproduce human-like decision-making behavior, we conducted a series of ablation studies. These ablations systematically removed or isolated different parts of the input, such as demographics, human-likeness prompts, prior dialogue components (arithmetic and memory), to identify what elements were essential for reproduction of bias and interaction behavior in a complex dialogue setting.

First, we removed demographic information from the inputs. The results remained mostly the same across all human-likeness levels. Accuracy stayed at 0.750 for HL1 and HL2, and HL3 showed slightly lower accuracy (0.500), similar correlation with human data (ρ = 0.771, p = .07), suggesting that demographics are not critical. Next, we tested a minimal setup with only the choice problem, no demographics, no dialogue, and no prompt. In this case, GPT-4.1 still showed biased choices, but the correlation with human patterns dropped (ρ = 0.257), indicating the model was biased but not in the same way as humans. We then tested only the memory component of the prior dialogue, where participants were asked to remember specific details. This condition improved alignment significantly, especially for GPT-4.1-mini under HL2 (accuracy = 0.833, ρ = 0.771, p = .07), showing that memory plays an important role in load-bias interaction. In contrast, when we kept only the arithmetic comparison component and removed memory cues, the models showed high accuracy but no meaningful correlation with human behavior (e.g., GPT-4.1: ρ = −0.086, p = .87). This suggests that arithmetic reasoning alone is not enough to model bias interaction. Overall, these findings show that while LLMs can reproduce basic bias effects, modeling human-like responses under cognitive load requires contextual elements, especially memory cues, in the dialogue.

To further examine whether LLMs rely on human responses in the prior dialogue when predicting individual human choices, we conducted a human response perturbation analysis. Specifically, we replaced the human responses in the chat transcripts (Section 4.1) with randomly generated text, while keeping all other aspects of the prompt unchanged.

We observed that the accuracy of the LLMs' predictions changed in fractions. The Figure 2 shows the deviation of individual level accuracy taking GPT4.1 accuracy as a reference. In the Figure 2, *gpt4_1_blrp* stands for GPT-4.1 baseline experiment with human

Table 8: Ablation study results showing the effect of removing or isolating different components (demographics, memory, and arithmetic components of prior dialogue) on model accuracy and alignment with human bias interactions. Correlation values (Pearson and Spearman) indicate how well the LLMs reproduce the directional change in bias under prior dialogue complexity.

Condition	Model (HL)	Accuracy	Spearman (ρ)
Demographics removed	GPT-4.1 (HL1)	0.750	0.257 (0.623)
	GPT-4.1 (HL2)	0.750	0.429 (0.397)
	GPT-4.1 (HL3)	0.500	**0.771 (0.072)**
Choice Problem Only	GPT-4.1	0.750	0.257 (0.623)
	GPT-4.1-mini	0.333	-0.600 (0.208)
Memory Component Followed by Choice Problem	GPT-4.1	0.750	0.429 (0.397)
	GPT-4.1-mini	0.833	**0.771 (0.072)**
Arithmetic Component Only Followed by Choice Problem	GPT-4.1	0.750	-0.086 (0.872)
	GPT-4.1-mini	0.750	0.029 (0.957)

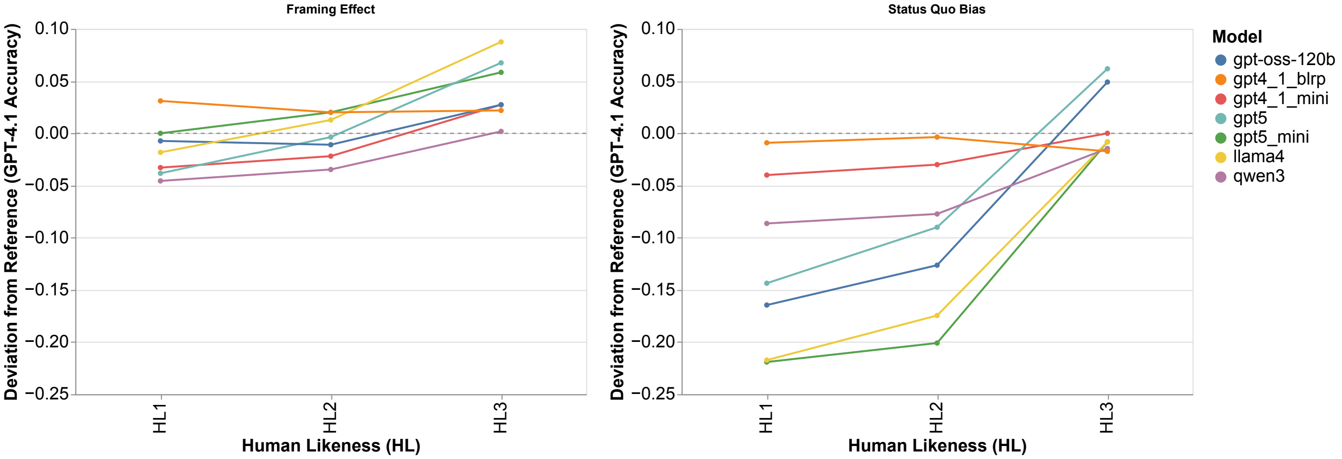

Figure 2: Deviation in accuracy from GPT-4.1 reference model across Human-Likeness (HL1, HL2, HL3) levels for each LLM. Positive values indicate higher accuracy than GPT-4.1; negative values indicate lower accuracy. gpt4_1_blrp stands for GPT-4.1 baseline experiment with human response perturbation.

response perturbation. For both Framing and Status quo, this trend line is closer to zero. This suggests that the models are primarily leveraging the structure of the prior dialogue, rather than fine-grained cues from human responses, to make their predictions.

5 Discussion

Our study investigated whether large language models (LLMs) can predict biased human decision-making in conversational settings. Addressing **RQ1**, biases such as the Framing and Status Quo effects were reproduced in human-conversational agent interactions, and their effects varied across choice problems. This extends inferring cognitive bias using simple cognitive tasks in traditional experiments into the conversational setting. This in-turn acting as a baseline for subsequent investigation.

RQ2 explored the role of complex prior dialogue. We found that increased dialogue complexity selectively increased susceptibility to the Framing effect, aligning with prior findings on working memory capacity and Framing [7, 57]. However, dialogue complexity did not affect Status Quo bias. Empirical work linking cognitive load to Status Quo bias is scarce; the bias is more commonly attributed to irrational emotional attachment [34]. Although choice overload has been shown to favor the Status Quo effect under cognitive load [16], we suspect that the simplicity of our binary choice task (e.g., College A vs. College B) limited such effects. This interpretation remains speculative and warrants further investigation using more sophisticated designs. Overall, our findings help address a gap identified in prior work (Section 2.1), where cognitive biases were largely studied in isolation, neglecting the role of conversational cognitive load, and suggest an empirical framework adaptable to other biases.

RQ3 investigated if simulated responses of LLM-agents can predict individual human decisions conditioned on limited prior dialogue and demographic information. The results are mixed; LLMs predicted individual human choices more accurately when conversational context accompanied demographic information. However, the degree of improvement varied by task and was highly dependent on the choice problem. In addition, ablation and perturbation analyses revealed that human-like responses under cognitive load depend on contextual cues, especially memory elements, rather than human utterances in dialogue.

While RQ3 investigated how closely the simulated responses aligned with human counterparts, **RQ4** examined whether simulated responses collectively exhibited biased behavior observed in human experiments. Results showed that models reproduced many human-like bias patterns (HL1 & HL2). However, sensitivity to cognitive load remained limited unless explicitly prompted (HL3 in Framing effect experiments). Moreover, explicit bias prompting (HL3 in Status quo bias experiments) led to overestimation, producing false positives where humans showed no bias. Alignment overfitting in language models can cause disproportionate adaptation to prompts, leading to over-interpretation of implied instructions [35, 58]. In our study, this appeared as increased sensitivity to HL3, inflating Status quo bias rather than reflecting human decision-making accurately.

5.1 Implications for LLM Simulation in HCI

Our findings contribute to ongoing discussions on whether LLMs can reproduce human cognitive biases in a conversational setting to be leveraged by LLM simulation. While recent work has suggested that LLMs may behave more rationally than humans [32], our study under HL1 and HL2 prompting conditions shows that LLMs can accurately reproduce biases. However, this alone does not confirm that LLMs are simulating human cognitive processes. It remains possible that these models are simply matching patterns based on learned statistical associations, especially given the widespread use of these choice problems in existing datasets. If the biased behavior observed in LLMs arises from statistical pattern matching rather than emulating underlying cognitive mechanisms, their use as human agents for behavioral simulation can be limited and sometimes misleading. In these cases, LLMs risk producing superficial or inaccurate representations of human decision-making, limiting their reliability for simulation.

Cognitive biases often interact with contextual factors such as cognitive load, as predicted by dual-process theory [28]. However, this interaction is not consistent across all types of biases. In our experiment focusing on the Framing, we found positive or negative change in the direction of effect size due to cognitive load (Table 2). However, in Status quo bias, we found that increased prior dialogue complexity consistently did not affect observed bias across all choice problems in human participants. In our LLM simulation experiment, most cases (Refer Table 7) in HL1 and HL2, where the models were instructed to act like humans, but without explicit reference to cognitive biases, LLMs failed to align with human behavior under load-bias interactions captured as correlation (ρ). In contrast, GPT4.1 in HL1 and HL2 showed 60% correlation, and under HL3 showed 77%, showing its alignment. Although the correlation is marginally significant, the findings leave an interesting note. The interactions converging toward human behavior suggest that LLMs may be capable of simulating context-sensitive bias patterns when explicitly prompted. However, this is only preliminary evidence, leaving open the possibility that LLM-generated behavior could more closely resemble human patterns and offer potential for more realistic behavioral simulation. Further investigation into a broader range of biases is required for stronger generalization.

5.2 Implications for Conversational AI

Recent studies show that LLMs are increasingly used as proxies for human participants. They are widely explored in different domains, such as social science, market analysis, and behavioral research [9, 22]. There is ongoing discussion in HCI about using LLMs as human proxies in empirical research [25]. This work evaluates how accurately LLM-generated data represents human decision-making in conversational settings. We compare real human choices with simulated choices under the same conversational conditions. Through this comparison, we contribute to recent research on LLM-based simulation by proposing a methodological approach for dialogue simulations driven by LLMs. Although the preliminary results were mixed, they suggest that LLMs have the capacity to predict biased patterns in human decision-making in a conversational setting. This suggests that LLMs may be useful proxies for modeling group-level user behavior in the design and evaluation of conversational agents. Using LLMs in this way can help researchers and practitioners estimate how typical users may respond to different agents' utterances, system behaviors, or dialogue flows. This can reduce the need for costly or complex user studies. As a result, LLMs may provide a scalable tool for A/B testing. However, future work is needed to identify which aspects of decision-making allow LLM proxies to be reliably used in practice.

From an application perspective, these findings have direct implications for the design and deployment of LLM-based conversational agents in interactive decision-making settings. The robust reproduction of classical biases suggests that decision-making facilitated by conversational AI can also result in the same cognitive biases as non-conversational interfaces. This highlights the need for LLM-assisted, bias-aware adaptive mechanisms. These mechanisms can be used to detect when users are especially susceptible to bias by leveraging the factors from prior dialogue. They can then adapt the dialogue accordingly and respond with a clearer and transparent presentation of alternatives. Because the presentation of alternatives invariably influences the decision-maker, and therefore, even a random or 'unthoughtful' presentation creates an impact (whether intended or not) [27]. This has led to terms like 'choice architecture' or 'nudging' being used to characterize the presentation of alternatives in decision scenarios [51]. This has further raised ethical questions, especially their impact on the user's autonomy [36]. Therefore, bias-aware adaptive systems could improve user outcomes in domains such as digital commerce, healthcare decision aids, or public service chatbots, where users routinely face complex decisions. Conversely, this predictive power can be used maliciously. Recent discussions on hypernudging [61] emphasize the ethical risks of exploiting cognitive biases via personal data. Our findings extend this concern by showing that influence does not necessarily require explicit personal information: LLMs can infer and predict the susceptibility of a human to biases from prior dialogue patterns and demographic cues alone. This raises new ethical challenges, calling for a critical examination of how adaptive conversational systems use dialogue history and context to influence human behavior.

Limitations

This study investigates Framing and Status Quo biases, providing a tightly controlled and replicable foundation for conversational agents and cognitive bias research. However, because different cognitive biases (e.g., anchoring, confirmation) may operate through distinct psychological mechanisms, these findings should not be assumed to generalize to other biases without further targeted research.

This work does not aim to characterize sources of dialogue complexity. Instead, we focus on settings where such complexity is already present and examine how the resulting cognitive load influences susceptibility to bias. Prior work has discussed dialogue strategies (e.g., open vs. closed-ended questions, conservative vs. non-conservative strategies, and goal alignment) as contributors to dialogue complexity. A detailed analysis of these strategies and dialogue naturalness is beyond the scope of this paper.

The study relies on self-reported (NASA-TLX) and behavioral (recall accuracy, response time) measures of cognitive load. While widely accepted, these do not capture real-time load. Future work could incorporate physiological measures (e.g., eye tracking, EEG, pupil dilation; [23]) for more dynamic assessment. Moreover, using dialogue complexity as a proxy for cognitive load may introduce unintended effects not captured by NASA-TLX, such as boredom, fatigue, distraction, or emotional states, which future studies could explicitly address.

Our LLM experiments were conducted using OpenAI's GPT-4.1 and GPT-5 families and selected open-source models. Although the approach is model-agnostic and extendable to other LLMs (e.g., Claude, Gemini, Mistral), broader comparisons are left to future work. This paper includes a minimal cross-generation and open-source comparison, and the released modular codebase supports replication with alternative models. Additionally, the human-likeness prompts used were basic; richer prompting strategies remain for future exploration.

Despite safeguards for data quality (memory recall tasks, attention checks, and response-time analysis), some data contamination remains possible. We found no clear evidence of LLM-assisted responses in human experiments, and Appendix G details our validation procedures. Model-side contamination is unlikely, as the LLMs used had a September 2024 training cutoff, while data were collected in 2025.

Although focused on dialogue, this approach generalizes to other settings, such as visual interfaces, where framed or status quo options induce bias. Nonetheless, dialogue remains a natural and effective paradigm for studying interactions between cognitive load and bias in human–agent interaction.

Statement of Ethics

All human-subject data were collected under ethical approval, anonymized prior to analysis, and handled in accordance with data protection guidelines; no personally identifiable demographic information, such as participant names or Prolific IDs, was given as input to the LLMs. This study was approved by the Institutional Review Board (IRB) and adheres to the ethical guidelines established by the institution. Details have been elided for anonymous review, can be provided on request. *(HREC-LS): LS-LR-24-278-Pilli-Nallur and LS-C-25-001-Pilli-Nallur.*

Acknowledgments

This work was conducted with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. This work was partially supported by the RE-ROUTE project, which has received funding from the European Union's Horizon Europe Marie Sklodowska-Curie Actions (MSCA), Staff Exchanges under Grant Agreement No. 101086343. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

GenAI Usage Disclosure

Tools such as Grammarly [21] and ChatGPT [37] were used solely for grammar checking and text polishing. No part of the conceptualization, experimental design, analysis, or interpretation was generated by any large language model. All substantive research contributions were produced entirely by the authors. In addition, GitHub Copilot [20] was used during code development to assist with routine programming tasks, such as syntax suggestions and code completion. All code was thoroughly reviewed, verified, and validated by the authors to ensure correctness and integrity.

References

[1] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*. PMLR, 337–371.

[2] Hugues Ali Mehenni, Sofiya Kobylyanskaya, Ioana Vasilescu, and Laurence Devillers. 2021. Nudges with Conversational Agents and Social Robots: A First Experiment with Children at a Primary School. In *Conversational Dialogue Systems for the Next Decade*, Luis Fernando D'Haro, Zoraida Callejas, and Satoshi Nakamura (Eds.). Springer, Singapore, 257–270. doi:10.1007/978-981-15-8395-7_19

[3] Avani Aravind, Sabyasachee Mishra, and Matt Meservy. 2024. Nudging towards sustainable urban mobility: Exploring behavioral interventions for promoting public transit. *Transportation Research Part D: Transport and Environment* 129 (2024), 104130.

[4] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[5] Jennifer E Arnold. 1998. *Reference form and discourse patterns.* Stanford University.

[6] Marcel Binz and Eric Schulz. 2024. Turning large language models into cognitive models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=eiC4BKypf1

[7] Boris Bogdanov, Jonathan Corbin, Sabina Dobreva, Todd McElroy, and Nikolay R Rachev. 2023. Working memory capacity and the risky-choice framing effect: A preregistered replication and extension of Corbin et al.(2010). *Judgment and Decision Making* 18 (2023), e39.

[8] Florian Brachten, Felix Brünker, Nicholas RJ Frick, Björn Ross, and Stefan Stieglitz. 2020. On the ability of virtual agents to decrease cognitive load: an experimental study. *Information Systems and e-Business Management* 18, 2 (2020), 187–207.

[9] James Brand, Ayelet Israeli, and Donald Ngwe. 2023. Using LLMs for market research. *Harvard business school marketing unit working paper* 23-062 (2023).

[10] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3290605.3300733

[11] Ping Chen. 1985. Discourse-T. Givón (ed.), Topic continuity in discourse: A quantitative cross-language study.(Typological Studies in Language, vol. 3.) Amsterdam and Philadelphia: John Benjamins, 1983. Pp. 492. *Language in Society* 14, 3 (1985), 410–414.

[12] Cary Deck and Salar Jahedi. 2015. The effect of cognitive load on economic decision making: A survey and new experiments. *European Economic Review* 78 (2015), 97–119.

[13] Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. Observations on typing from 136 million keystrokes. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[14] Mateusz Dubiel, Anastasia Sergeeva, and Luis A. Leiva. 2024. Impact of Voice Fidelity on Decision Making: A Potential Dark Pattern?. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. ACM, Greenville SC USA, 181–194. doi:10.1145/3640543.3645202

[15] Jessica Maria Echterhoff, Matin Yarmand, and Julian McAuley. 2022. AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–9. doi:10.1145/3491102.3517443

[16] Scott Eidelman and Christian S. Crandall. 2012. Bias in Favor of the Status Quo. *Social and Personality Psychology Compass* 6, 3 (2012), 270–281. doi:10.1111/j.1751-9004.2012.00427.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-9004.2012.00427.x.

[17] Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis (Eds.). Association for Computational Linguistics, Saarbrücken, Germany, 207–219. doi:10.18653/v1/W17-5526

[18] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[19] Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68, 1 (1998), 1–76.

[20] GitHub, Inc. 2025. GitHub Copilot Documentation. https://docs.github.com/en/copilot. Accessed: 2025-12-27.

[21] Grammarly Inc. 2025. Grammarly. https://www.grammarly.com. Accessed: 2025-07-30.

[22] Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. AI and the transformation of social science research. *Science* 380, 6650 (2023), 1108–1109.

[23] Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (Copenhagen, Denmark) *(UbiComp '10)*. Association for Computing Machinery, New York, NY, USA, 301–310. doi:10.1145/1864349.1864395

[24] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[25] Angel Hsing-Chi Hwang, Michael S Bernstein, S Shyam Sundar, Renwen Zhang, Manoel Horta Ribeiro, Yingdan Lu, Serina Chang, Tongshuang Wu, Aimei Yang, Dmitri Williams, et al. 2025. Human Subjects Research in the Age of Generative AI: Opportunities and Challenges of Applying LLM-Simulated Data to HCI Studies. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[26] Kaixin Ji, Sachin Pathiyan Cherumanal, Johanne R. Trippas, Danula Hettiachchi, Flora D. Salim, Falk Scholer, and Damiano Spina. 2024. Towards Detecting and Mitigating Cognitive Bias in Spoken Conversational Search. In *26th International Conference on Mobile Human-Computer Interaction*. ACM, Melbourne VIC Australia, 1–10. doi:10.1145/3640471.3680245

[27] Eric J. Johnson, Suzanne B. Shu, Benedict G. C. Dellaert, Craig Fox, Daniel G. Goldstein, Gerald Häubl, Richard P. Larrick, John W. Payne, Ellen Peters, David Schkade, Brian Wansink, and Elke U. Weber. 2012. Beyond nudges: Tools of a choice architecture. *Mark Lett* 23, 2 (June 2012), 487–504. doi:10.1007/s11002-012-9186-1

[28] Daniel Kahneman. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, NY, US. Pages: 499.

[29] Natalia Kalashnikova, Ioana Vasilescu, and Laurence Devillers. 2024. Linguistic Nudges and Verbal Interaction with Robots, Smart-Speakers, and Humans. (2024).

[30] Adwait Khare, Tilottama G Chowdhury, and Jeremy Morgan. 2021. Maximizers and Satisficers: Can't choose and Can't reject. *Journal of Business Research* 135 (2021), 731–748.

[31] Yi Kuang, Yuan-Na Huang, and Shu Li. 2023. A framing effect of intertemporal and spatial choice. *Quarterly Journal of Experimental Psychology* 76, 6 (2023), 1298–1320.

[32] Ryan Liu, 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 *Proceedings of the International Conference on Learning Representations (ICLR)*.

[33] Simon Malberg, Roman Poletukhin, Carolin M. Schuster, and Georg Groh. 2025. A Comprehensive Evaluation of Cognitive Biases in LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar (Eds.). Association for Computational Linguistics, Albuquerque, USA, 578–613. doi:10.18653/v1/2025.nlp4dh-1.50

[34] Yusufcan Masatlioglu and Efe A. Ok. 2005. Rational choice with status quo bias. *Journal of Economic Theory* 121, 1 (March 2005), 1–29. doi:10.1016/j.jet.2004.03.007

[35] Erik Miehling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M Daly, Kush R Varshney, Eitan Farchi, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, Miao Liu, et al. 2025. Evaluating the prompt steerability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 7874–7900.

[36] Vivek Nallur, Karen Renaud, and Aleksei Gudkov. 2025. Nudging Using Autonomous Agents: Risks and Ethical Considerations. In *Multi-Agent Systems*, Rem Collier, Alessandro Ricci, Vivek Nallur, Samuele Burattini, and Andrea Omicini (Eds.). Springer Nature Switzerland, Cham, 283–296.

[37] OpenAI. 2024. ChatGPT-4o [Computer software]. https://openai.com/chatgpt. Accessed: 2025-08-05.

[38] Bhavna Pancholi, Mark Dunne, and Richard Armstrong. 2009. Sample size estimation and statistical power analyses. 16 (11 2009).

[39] Vinoth Pandian Sermuga Pandian and Sarah Suleri. 2020. NASA-TLX Web App: An Online Tool to Analyse Subjective Workload. http://arxiv.org/abs/2001.09963 arXiv:2001.09963 [cs].

[40] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[41] Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109* (2024).

[42] Stephen Pilli. 2023. Exploring conversational agents as an effective tool for measuring cognitive biases in decision-making. In *2023 10th International Conference on Behavioural and Social Computing (BESC)*. IEEE, 1–5.

[43] Ganna Pogrebna, Karen Renaud, and Marina Kovaleva. 2025. *The big bad bias book.*

[44] Prolific. 2024. Prolific. https://www.prolific.com. First released in 2014. Current version accessed in September 2025. London, UK..

[45] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (April 2020), 8689–8696. doi:10.1609/aaai.v34i05.6394 Number: 05.

[46] William Samuelson and Richard Zeckhauser. 1988. Status quo bias in decision making. *J Risk Uncertainty* 1, 1 (March 1988), 7–59. doi:10.1007/BF00055564

[47] Johanna Schmidhuber, Stephan Schlögl, and Christian Ploder. 2021. Cognitive Load and Productivity Implications in Human-Chatbot Interaction. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*. 1–6. doi:10.1109/ICHMS53169.2021.9582445

[48] Herbert A. Simon. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* 69, 1 (1955), 99–118. doi:10.2307/1884852 Publisher: Oxford University Press.

[49] Streamlit. 2019. Streamlit: A Faster Way to Build and Share Data Apps. Available at https://streamlit.io/. Accessed: 2025-02-27.

[50] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.

[51] Richard H. Thaler, Cass R. Sunstein, and John P. Balz. 2010. Choice Architecture. doi:10.2139/ssrn.1583509

[52] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. 185 (1974).

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

[54] Stefan Ultes and Wolfgang Maier. 2020. On the Complexity in Task-oriented Spoken Dialogue Systems. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–4.

[55] Teun Adrianus Van Dijk, Walter Kintsch, et al. 1983. Strategies of discourse comprehension. (1983).

[56] Xiao Tian Wang. 1996. Framing effects: Dynamics and task domains. *Organizational behavior and human decision processes* 68, 2 (1996), 145–157.

[57] Paul Whitney, Christa A Rinehart, and John M Hinson. 2008. Framing effects under cognitive load: The role of working memory in risky decisions. *Psychonomic bulletin & review* 15, 6 (2008), 1179–1184.

[58] Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2024. Fundamental limitations of alignment in large language models. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 2176, 34 pages.

[59] Qinyu Xiao, Emma Lam, Muhrajan Piara, and Gilad Feldman. 2021. Revisiting status quo bias: Replication of Samuelson and Zeckhauser (1988). *Meta-Psychology* 5 (Feb. 2021). doi:10.15626/MP.2020.2470

[60] Yusuke Yamamoto. 2024. Suggestive answers strategy in human-chatbot interaction: a route to engaged critical decision making. *Frontiers in Psychology* 15 (March 2024), 1382234. doi:10.3389/fpsyg.2024.1382234

[61] Karen Yeung. 2019. 'Hypernudge': Big Data as a mode of regulation by design. In *The social power of algorithms*. Routledge, 118–136.

[62] Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. 2025. A Survey on Recent Advances in LLM-Based Multi-turn Dialogue Systems. *ACM Comput. Surv.* 58, 6, Article 148 (Dec. 2025), 38 pages. doi:10.1145/3771090

[63] Lance Ying, Katherine M Collins, Lionel Wong, Ilia Sucholutsky, Ryan Liu, Adrian Weller, Tianmin Shu, Thomas L Griffiths, and Joshua B Tenenbaum. 2025. On benchmarking human-like intelligence in machines. *arXiv preprint arXiv:2502.20502* (2025).

A Choice Problems

A.1 Framing Choice Problems

Table 9: Choice problems, types, and conditions (framed and alternatively framed) and their respective alternatives. Previous studies indicate that individuals are biased towards the alternatives highlighted in bold.

Choice Problem Type	Choice Problem	Choice Problem Framing Condition	Alternatives
Risky Choice Framing (RCF)	Imagine that after a serious traffic accident, 100 people are stranded in a tunnel. As a Public Transportation Officer choose between two plans.	Saved	***If plan A is adopted, 25 people will be saved.*** If plan B is adopted, there is a 1/4 chance of saving all 100 people and a 3/4 chance of not saving anyone.
		Lost	If plan A is adopted, 75 people will die. ***If plan B is adopted, there is a 1/4 chance that no people will die and a 3/4 chance that all 100 people will die.***
Attribute Framing (ATF)	Suppose you are planning to dine out. Two restaurants are available. The only way to go to both restaurants from your home is by public transportation. Which one would you prefer? Note: The average speed of a bus is approximately 0.41 mi/min (25 mph).	Space	***The first restaurant is approximately 5 miles by bus from your home, and the restaurant's star rating is 6/10.*** The second restaurant is approximately 9 miles by bus from your home, and the restaurant's star rating is 7/10.
		Time	The first restaurant is approximately 12 min by bus from your home, and the restaurant's star rating is 6/10. T***he second restaurant is approximately 22 min by bus from your home, and the restaurant's star rating is 7/10.***
Goal Framing (GF)	Let's assume that you are travelling to a place 10 miles away, given that the two modes of transportation are available. If you need to choose between one of the two options shown below, which would you choose?	No Goal	***Personal car***. Public transport.
		With Goal	Note: It has been found that by switching from a 20-mile commute by car to public transport, an individual can reduce their annual CO2 emissions by around 9 kg per day, or more than 21,700 kg per year. Personal car. ***Public transport***.

A.2 Status Quo Choice Problems

A.2.1 Budget Allocation (SC1).

Neutral Condition.

The National Highway Safety Commission is deciding how to allocate its budget between two safety research programs:

- Improving automobile safety (bumpers, body, gas tank configuration, seat-belts), and
- Improving the safety of interstate highways (guard rails, grading, highway interchanges, and implementing selective reduced speed limits).

Since there is a ceiling on its total spending, it must choose between the options provided below. If you had to make this choice, which of the following will you choose?

- Allocate 60% to auto safety and 40% to highway safety
- Allocate 50% to auto safety and 50% to highway safety

Neutral Condition - Alternative Swapped.

The National Highway Safety Commission is deciding how to allocate its budget between two safety research programs:

- Improving automobile safety (bumpers, body, gas tank configuration, seat-belts), and
- Improving the safety of interstate highways (guard rails, grading, highway interchanges, and implementing selective reduced speed limits).

Since there is a ceiling on its total spending, it must choose between the options provided below. If you had to make this choice, which of the following will you choose?

- Allocate 50% to auto safety and 50% to highway safety.
- Allocate 60% to auto safety and 40% to highway safety.

Status Quo A - 60A40H .

The National Highway Safety Commission is deciding how to allocate its budget between two safety research programs:

- Improving automobile safety (bumpers, body, gas tank configuration, seat-belts)
- Improving the safety of interstate highways (guard rails, grading, highway interchanges, and implementing selective reduced speed limits).

Currently, the commission allocates approximately 60% of its funds to auto safety and 40% of its funds to highway safety. Since there is a ceiling on its total spending, it must choose between the options provided below. If you had to make this choice, which of the following will you choose?

- Maintain present budget amounts for the programs.
- Decrease auto program by 10% and raise highway program by like amount.

Status Quo B - 50A50H .

The National Highway Safety Commission is deciding how to allocate its budget between two safety research programs:

- Improving automobile safety (bumpers, body, gas tank configuration, seat-belts)
- Improving the safety of interstate highways (guard rails, grading, highway interchanges, and implementing selective reduced speed limits).

Currently, the commission allocates approximately 50% of its funds to auto safety and 50% of its funds to highway safety. Since there is a ceiling on its total spending, it must choose between the options provided below. If you had to make this choice, which of the following will you choose?

- Maintain present budget amounts for the programs.
- Increase auto program by 10% and lower highway program by like amount.

A.2.2 Investment Decision Making (SC3).

Neutral Condition.

You are a serious reader of the financial pages but until recently have had few funds to invest. That is when you inherited a large sum of money from your great uncle. You are considering different portfolios. Your choices are:

- Invest in moderate-risk Company A. Over a year's time, the stock has .5 chance of increasing 30% in value, a .2 chance of being unchanged, and a .3 chance of declining 20% in value.
- Invest in high-risk Company B. Over a year's time, the stock has a .4 chance of doubling in value, a .3 chance of being unchanged, and a .3 chance of declining 40% in value.

Neutral Condition - Alternative Swapped.

You are a serious reader of the financial pages but until recently have had few funds to invest. That is when you inherited a large sum of money from your great uncle. You are considering different portfolios. Your choices are:

- Invest in high-risk Company B. Over a year's time, the stock has a .4 chance of doubling in value, a .3 chance of being unchanged, and a .3 chance of declining 40% in value.
- Invest in moderate-risk Company A. Over a year's time, the stock has .5 chance of increasing 30% in value, a .2 chance of being unchanged, and a .3 chance of declining 20% in value.

Status Quo A - Moderate Risk.

You are a serious reader of the financial pages but until recently have had few funds to invest. That is when you inherited a portfolio of cash and securities from your great uncle. A significant portion of this portfolio is invested in moderate-risk Company A. You are deliberating whether to leave the portfolio intact or change it by investing in other securities. (The tax and broker commission consequences of any change are insignificant.) Your choices are:

- Retain the investment in moderate-risk Company A. Over a year's time, the stock has .5 chance of increasing 30% in value, a .2 chance of being unchanged, and a .3 chance of declining 20% in value.
- Invest in high-risk Company B. Over a year's time, the stock has a .4 chance of doubling in value, a .3 chance of being unchanged, and a .3 chance of declining 40% in value.

Status Quo B - High Risk.

You are a serious reader of the financial pages but until recently have had few funds to invest. That is when you inherited a portfolio of cash and securities from your great uncle. A significant portion of this portfolio is invested in high-risk Company B. You are deliberating whether to leave the portfolio intact or change it by investing in other securities. (The tax and broker commission consequences of any change are insignificant.) Your choices are:

- Retain the investment in high-risk Company B. Over a year's time, the stock has a .4 chance of doubling in value, a .3 chance of being unchanged, and a .3 chance of declining 40% in value.
- Invest in moderate-risk Company A. Over a year's time, the stock has a .5 chance of increasing 30% in value, a .2 chance of being unchanged, and a .3 chance of declining 20% in value.

A.2.3 College Jobs (SC4).

Neutral Condition.

Having just completed your graduate degree, you have two offers of teaching jobs in hand. When evaluating teaching job offers, people typically consider the salary, the reputation of the school, the location of the school, and the likelihood of getting tenure (tenure is permanent job contract that can only be terminated for cause or under extraordinary circumstances). Your choices are:

- College A: east coast, very prestigious school, high salary, fair chance of tenure.
- College B: west coast, low prestige school, high salary, good chance of tenure.

Neutral Condition - Alternative Swapped.

Having just completed your graduate degree, you have two offers of teaching jobs in hand. When evaluating teaching job offers, people typically consider the salary, the reputation of the school, the location of the school, and the likelihood of getting tenure (tenure is permanent job contract that can only be terminated for cause or under extraordinary circumstances). Your choices are:

- College B: west coast, low prestige school, high salary, good chance of tenure.
- College A: east coast, very prestigious school, high salary, fair chance of tenure.

Status Quo A - College A.

You are currently an assistant professor at College A in the east coast. Recently, you have been approached by colleague at other university with job opportunity. When evaluating teaching job offers, people typically consider the salary, the reputation of the school, the location of the school, and the likelihood of getting tenure (tenure is permanent job contract that can only be terminated for cause or under extraordinary circumstances). Your choices are:

- Remain at College A: east coast, very prestigious school, high salary, fair chance of tenure.
- Move to College B: west coast, low prestige school, high salary, good chance of tenure.

Status Quo B - College B.

You are currently an assistant professor at College B in the west coast. Recently, you have been approached by colleague at other university with job opportunity. When evaluating teaching job offers, people typically consider the salary, the reputation of the school, the location of the school, and the likelihood of getting tenure (tenure is permanent job contract that can only be terminated for cause or under extraordinary circumstances). Your choices are:

- Remain at College B: west coast, low prestige school, high salary, good chance of tenure.
- Move to College A: east coast, very prestigious school, high salary, fair chance of tenure.

A.3 Textual Adjustments to Original Choice Problems

We made a few syntactic adjustments to the choice problems to ensure consistency across different experimental conditions. Below, we outline the key modifications made to the phrasing, structure, and presentation of the scenarios.

(1) **Risky Choice Framing:** Depending on the experimental condition, the alternatives are framed in terms of either lives saved or lives lost. The original choice problem remains unchanged, with the only modification being the replacement of "public authorities" with "Public Transportation Officer".

(2) **Attribute Framing:** The adapted attribute framing choice problem is largely unmodified. The minor modifications that we made are to make sure the presentation of alternatives is conversational and user interface-friendly. Adapting to the demographics, we change from kilometers to miles in the choice problem. The original alternatives do not use the terms "first" and "second", instead, they use a numbered list as they are tailored for digital user interfaces. We modified *"A. The restaurant..."* to *"The first restaurant..."*. This is suitable for a conversational user interface. Further, we have replaced "drive" with "public transportation" and "car" with "bus". The star rating changed from visual ★ ★ ★ ★ ★ ★ ★ ☆ ☆☆ to text "7/10".

(3) **Goal Framing:** We did not perform any substantial modification of the original choice problem; we only changed the units from pounds to kilograms while presenting the CO2 information. Further, to simplify the alternatives, we have excluded the multi-modal transportation option from the list of original alternatives.

(4) **Budget Allocation:** No changes were made.

(5) **Investment Decision Making:** No changes were made.

(6) **College Jobs:** In the neutral condition, we reduced the number of teaching job offers from four to two. Similarly, in the Status Quo condition, we changed the text "*colleagues at other universities with job opportunities*" to "*colleague at other university with job opportunity*".

In all the choice problems, we moved from an ordered list presentation to bullet points. The replication study for status quo by Xiao et al. [59] included a method to assess whether participants understood the choice problem. Participants were first shown the scenario and asked various related questions before being presented with the decision alternatives. To maintain conversational fluidity, we asked the comprehension questions later in the survey (as *Choice Problem Attention Check*), and not during the chatbot's interaction. Accuracy on the choice problem Attention Check was used as a filtering criterion. Additionally, we asked participants whether they had encountered the choice problem before, recorded as *Choice Problem Familiarity*. To mitigate potential learning effects, this variable was also used to exclude data from participants with prior exposure.

B Prior Dialogue

B.1 Simple Dialogue

Table 10: Yes/No questions of preference elicitation tasks, designed using a less conservative dialogue strategy, across five different domains of the SGD Dataset. Please enter "I don't know" as an attention check.

Domain	Attribute	Yes/No Questions
Real Estate	Budget	Do you have a specific budget for the home?
	Location	Are you looking for a home in a specific location?
	Number of Bedrooms	Do you need more than 3 bedrooms?
	Number of Bathrooms	Is having 2 or more bathrooms important to you?
	Type of Home (Apartment, House, etc.)	Are you looking specifically for a detached house? Please enter "I don't know" only.
	Size (Square Footage)	Do you prefer homes larger than 2000 square feet?
Music	Genre Preference	Do you like listening to pop music?
	Language of Lyrics	Do you prefer music with lyrics in English?
	Live Performances	Are you interested in live music performances?
	Instruments Focused	Do you enjoy instrumental music?
	Artist-Specific	Do you like music from specific artists? Please enter "I don't know" only.
	Era (e.g., 80s, 90s)	Do you prefer music from the 90s?
Movies	Genre Preference	Do you like action movies?
	Language Preference	Do you prefer movies in English?
	Director/Actor Specific	Are you a fan of movies by a specific director or actor?
	Streaming Service Preference	Do you prefer to watch movies on Netflix?
	High Definition (HD)	Do you prefer movies in high definition? Please enter "I don't know" only.
	Subtitles	Do you need subtitles when watching movies?
Calendar	Daily Reminders	Do you need daily reminders?
	Meeting Notifications	Do you want notifications for upcoming meetings?
	Recurring Events	Do you often schedule recurring events?
	Shared Calendars	Do you use shared calendars with others?
	Event Location	Is specifying the location of events important to you? Please enter "I don't know" only.
	Time Zone Adjustments	Do you need automatic time zone adjustments for events?
Banks	Online Banking	Do you prefer to use online banking?
	Mobile Banking	Do you want to manage your accounts through mobile banking apps?
	Bank Branch Proximity	Is it important that a bank branch is close to your home or work?
	ATM Availability	Is having nearby ATMs a priority for you?
	Interest Rates	Are competitive interest rates important to you? Please enter "I don't know" only.
	Fee-Free Accounts	Do you prefer accounts with no fees?
Apps	Security	Do you prefer end-to-end encryption (security) for enhanced privacy?
	Cross-Platform	Is cross-platform compatibility (iOS, Android, desktop) important to you?
	File Sharing	Do you value the ability to send multimedia files like photos, videos, and documents?
	Group Messaging	Is group messaging and the ability to create large groups important for your communication needs?
	Calling Features	Do you prefer apps with voice and video calling features? Please enter "I don't know" only.
	Customisability	Are customisable notifications and settings important to you?

B.2 Complex Dialogue

Domain	Attribute			Questions
Home Property	**Number of Bedrooms**	**Size (Sq ft.)**	**Property Reviews**	
1st	Three Rooms	2000	Four star	In the following scenario choose from various property recommendations. The first property has three bedrooms, 2000 square feet, and a 4-star rating. The second property has twice the number of bedrooms and with the same size and rating. Which one do you prefer, and why? The third property has the same number of bedrooms as the second one but is half the size of the first one, with the same rating as the first. Which one do you prefer, and why? The fourth property has the same number of bedrooms as the second, the same size as the third, but one less star rating than the first. Which one do you prefer, and why? Remember the details of the fourth property. Specific information will be requested later.
2nd	Two times 1st	Same as first	Same as first	
3rd	Same as the second	Half of first	Same as first	
4th	Same as the second	Same as third	One star less than the first	
Music Artist	**Live Performances**	**Artist Remuneration for Show**	**Artist-Specific**	
1st	Three	2000 Units	Four star	In the following scenario choose from various artist recommendations. The first artist performs three live shows, is paid 2000 units per show, and has a 4-star rating. The second artist performs twice as many shows, with the same pay and rating. Which artist do you prefer, and why? The third artist performs the same number of shows as the second, earns half the pay of the first artist, but has the same rating as the first. Which artist do you prefer, and why? The fourth artist performs the same number of shows as the second, earns the same pay as the third, but has two stars less than the first artist. Which artist do you prefer, and why? Remember the details of the fourth artist. Specific information will be requested later.
2nd	2 times 1st	Same as first	Same as first	
3rd	Same as the second	Half of first	Same as first	
4th	Same as the second	Same as third	One star less than the first	

Table 11 continued from previous page

Domain	Attribute			Questions
Movies - Streaming Service	**Number of Parallel Devices**	**Library Size**	**Service Rating**	
1st	Three	2000 Movies	Four star	In the following scenario choose from various streaming service recommendations. The first streaming service supports 3 parallel devices, has a library of 2000 movies, and is rated 4 stars. The second service supports twice as many devices, with the same library size and rating. Which service do you prefer, and why? The third streaming service supports the same number of devices as the second, has half the library size of the first, but has the same rating as the first. Which service do you prefer, and why? The fourth streaming service supports the same number of devices as the second, has the same library size as the third, but is rated one star less than the first. Which service do you prefer, and why? Remember the details of the fourth service. Specific information will be requested later.
2nd	2 times 1st	Same as first	Same as first	
3rd	Same as the second	Half of first	Same as first	
4th	Same as the second	Same as third	One star less than the first	
Calendar App	**Calendar Syncing Across Devices**	**Managed Tasks Per Year**	**Event Privacy Rating**	
1st	3	2000	Four star	In the following scenario choose from Various Apps recommendations for calendar. The first calendar app can sync across three devices, manages 2000 tasks per year, and has a 4-star privacy rating. The second app syncs across two devices, manages the same number of tasks, and has the same privacy rating. Which app do you prefer, and why? The third app syncs across the same number of devices as the second app, but manages half as many tasks as the first app, with the same privacy rating as the first. Which app do you prefer, and why? The fourth app syncs across the same number of devices as the second, manages the same number of tasks as the third, but has one less star in privacy rating compared to the first. Which app do you prefer, and why? Remember the details of the fourth App. Specific information will be requested later.
2nd	2 times 1st	Same as first	Same as first	
3rd	Same as Second	Half of first	Same as first	
4th	Same as Second	Same as third	One star less than the first	

Table 11 continued from previous page

Domain	Attribute			Questions
Bank	**Bank Branch Proximity**	**Interest Rates**	**Fee-Free Accounts Rating**	
1st	3km	2000 units	Four	In the following scenario choose from various banks recommendations. The first bank is 3 km away, offers 2000 units of interest, and has a four-star rating for fee-free accounts.
2nd	2 times 1st	Same as first	Same as first	The second bank is twice as far away, offers the same amount of interest, and has the same fee-free account rating. Which bank would you prefer, and why?
3rd	Same as the second	Half of first	Same as first	The third bank is as far away as the second bank, offers half the amount of interest as the first bank, but has the same fee-free account rating as the first bank. Which bank would you prefer, and why?
4th	Same as the second	Same as third	One star less than the first	The fourth bank is as far away as the second bank, offers the same amount of interest as the third bank, but has one star less in fee-free account rating compared to the first bank. Which bank would you prefer, and why?
Messaging App	**Number of Simultaneous Devices**	**Messages Per Day**	**Security Rating**	
1st	3	2000	Four star	In the following scenario, choose from various messaging app recommendations. The first messaging app allows access on three devices, supports 2000 messages per day, and has a 4-star security rating.
2nd	2 times 1st	Same as first	Same as first	The second app allows access on twice as many devices, supports the same number of messages, and has the same security rating. Which app do you prefer, and why?
3rd	Same as the second	Half of first	Same as first	The third app allows access on the same number of devices as the second app, but supports half as many messages as the first app, with the same security rating as the first. Which app do you prefer, and why?
4th	Same as the second	Same as third	One star less than the first	The fourth app allows access on the same number of devices as the second, supports the same number of messages as the third, but has one less star in security rating compared to the first. Which app do you prefer, and why?

Table 11 continued from previous page

Domain	Attribute	Questions
		Remember the details of the fourth app, including the number of devices, messages per day, and its security rating. Specific information will be requested later.

Table 11: Complex Tasks for prior dialogue.

C A Priori Power Analysis

We conducted a power analysis using G*Power [18] to determine the required sample size, targeting a power of 0.80 to detect a medium effect size ($\omega = 0.3$). This choice reflects our focus on comparing effect sizes across conditions rather than on individual statistical significance. With $\alpha = 0.05$ and 1 degree of freedom [38], the required sample size was estimated to be 42 participants per condition. To ensure robustness, we recruited slightly more participants than required. For the Status quo experiments, which followed a 2 × 3 design with two dialogue complexity conditions (Simple vs. Complex) and three status quo conditions (Neutral, Status Quo A, Status Quo B), the minimal required sample size was 756 participants (42 per condition). For the Framing experiments, which followed a 2 × 2 design with two dialogue complexity conditions (Simple vs. Complex) and two framing conditions (Framed vs. Alternatively Framed), the minimal required sample size was 528 participants (44 per condition). Again, we recruited more than this minimum to strengthen the validity of our results.

C.1 Participant Recruitment, Compensation, and Pre-Registration

Participants were recruited through Prolific, a widely used platform known for ensuring data quality and participant reliability[44]. The estimated completion time for the survey was eight minutes, and participants were compensated according to Prolific's recommended minimum rate of $8 per hour. A total of 1648 participants were recruited, and measures were put in place to prevent duplicate participation. Additionally, demographic information for each participant was obtained from Prolific. The hypotheses for each experiment and experimental design were preregistered on the Open Science Framework to promote transparency. The studies were preregistered on the Open Science Framework (OSF). The preregistration for Framing study is archived at https://doi.org/10.17605/OSF.IO/DPR45, and the preregistration Status quo study is archived at https://doi.org/10.17605/OSF.IO/PSXVF.

C.2 Data Quality and Integrity

To ensure data integrity and control for familiarity bias, participants reported post-interaction whether they had previously encountered the choice problem and completed a domain-familiarity assessment. Attentiveness was evaluated via a memory-recall task based on the chatbot interaction, and participants were instructed to forgo external aids to preserve the validity of our cognitive-load manipulation. An automated system logged responses as JSON files, which were securely emailed to the authors and a public address to guarantee transparent data collection. The dataset for Framing effect study is available at https://doi.org/10.5281/zenodo.18218753, and the Status quo bias study is available at https://doi.org/10.5281/zenodo.16541481.

D Demographics

D.1 Framing Demographics

Table 12: Detailed Demographics Split by Choice Problem Type, Framing Condition, and Prior Discourse Conditions. All the Participants are From UK.

Prior Discourse Condition	Choice Problem Type	Choice Problem Condition	Sample Size (n)	Age (Mean)	Age (SD)	Sex (Levels)	Sex (counts)	Ethnicity (levels)	Ethnicity (counts)
No Load	Risky Choice Framing	Framing	45	41.6	12.9	Female Male	23 22	White Black Mixed	36 5 3
		Alternative Framing	44	42.8	12.1	Female Male	26 18	White Black Asian	34 4 3
	Attribute Framing	Framing	49	43.6	14.1	Female Male	27 22	White Mixed Asian	44 4 1
		Alternative Framing	47	46.4	12.3	Male Female	26 21	White Asian Black	43 1 1
	Goal Framing	Framing	44	44.2	13.6	Female Male	23 21	White Black Mixed	42 1 1
		Alternative Framing	44	41.9	14.9	Female Male	23 21	White Asian Black	35 3 3
Load	Risky Choice Framing	Framing	44	40.4	13.6	Female Male	26 18	White Black Asian	37 4 3
		Alternative Framing	44	41.9	14.8	Female Male	25 19	White Black Asian	33 10 1
	Attribute Framing	Framing	46	45.4	14.7	Female Male	23 23	White Asian	45 1
		Alternative Framing	53	44.3	12.3	Female Male	30 23	White Mixed Other	48 2 2
	Goal Framing	Framing	44	45.6	15	Male Female	23 21	White Other Asian	41 2 1
		Alternative Framing	44	37.2	13.5	Female Male	25 19	White Asian Black	36 3 2

D.2 Status Quo Demographics

Table 13: Demographic characteristics of participants across experimental conditions, split by prior dialogue condition (No Load vs. Load), choice problem (Budget Allocation, Investment Decision Making, College Jobs), and choice problem condition (Neutral, Status Quo A, Status Quo B). Reported variables include sample size (n), age (Mean, SD), country distribution (United Kingdom, United States, Ireland), and sex (Female, Male).

Prior Dialogue Condition	Choice Problem	Choice Problem Condition	n	Age		Country			Sex	
				Mean	SD	United Kingdom	United States	Ireland	Female	Male
No Load	BA	NEUT	60	42.5	13.6	53	6	1	32	28
		A	51	40.9	12.0	44	7	0	27	24
		B	73	43.0	13.4	49	22	2	39	34
	IDM	NEUT	51	35.5	12.6	33	18	0	21	30
		A	58	38.9	12.1	26	31	1	45	13
		B	54	43.7	11.6	47	7	0	10	44
	CJ	NEUT	76	45.8	14.7	60	16	0	35	41
		A	70	41.9	13.1	53	17	0	35	35
		B	51	37.5	12.9	38	13	0	26	25
Load	BA	NEUT	70	42.2	13.6	57	12	1	30	40
		A	59	39.7	14.6	48	10	1	29	30
		B	64	40.7	13.8	36	26	2	31	33
	IDM	NEUT	57	40.1	14.4	25	30	2	36	21
		A	51	40.9	11.4	26	25	0	36	15
		B	56	45.3	14.3	48	7	1	24	32
	CJ	NEUT	80	40.7	13.0	48	29	3	40	40
		A	70	42.0	12.5	57	13	0	33	37
		B	49	42.4	11.7	30	19	0	27	22

E Perceived Cognitive Load

We compared NASA-TLX scores and performance metrics between Simple and Complex dialogue conditions to confirm that complex prior dialogue results in cognitive load in chatbot interactions.

E.1 Framing

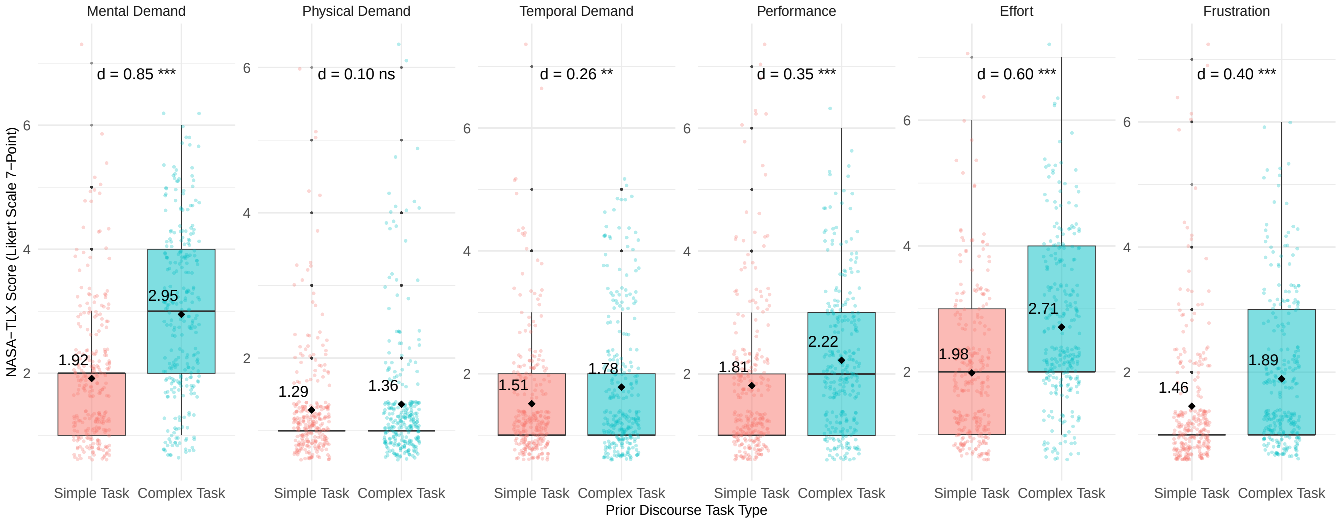

Figure 3: Boxplots, Effect Sizes, Significances ($* * *p < 0.001$, ns - no significance), and Means of NASA-TLX Scores for Simple vs. Complex Task Conditions.

We conducted a t-test to analyse differences in participants' NASA-TLX scores across each dimension when performing a Simple Vs Complex Dialogue. This analysis aimed to determine whether the perceived workload varied significantly based on task complexity. Figure 4 presents the effect sizes (Cohen's d) and significance levels, and box plots for NASA-TLX workload assessment dimensions.

The results show that Mental Demand had the largest differences between Simple and Complex Tasks across all choice problems. As shown in Figure 4 and supported by the data, the effect size was large (d=0.85, p<0.001), suggesting that participants experienced a much higher mental demand when performing the Complex Task. Effort also showed a significant difference with a medium effect size (d=0.60, p<0.001), indicating that participants required more effort under the Complex Task condition. Similarly, Frustration showed a small to medium effect size (d=0.40, p<0.001). For Performance, the effect was also statistically significant (d=0.35, p<0.001), which may indicate a perceived reduction in performance under the Complex Task condition. Temporal Demand showed a smaller yet significant effect (d=0.26, p<0.001). Physical Demand, on the other hand, did not show a statistically significant difference between task types (d=0.10, p=0.25), confirming that physical workload was not a major factor in this study's task design.

Moreover, a Linear mixed-effects models were used to assess whether task domains (with 6 levels) had a random effect on NASA-TLX Mental Demand scores. In both Simple and Complex Tasks, the estimated variance for the Domain as a random intercept was small (< 0.008), indicating limited between-domain variability. Most of the variation was attributed to individual differences (residual variance > 1.8). These findings suggest that perceived mental demand was largely consistent across task domains. Overall, Mental Demand, as highlighted in Figure 4, was the most affected by task complexity.

Our survey captured participants' recall performance on the memory component of the Complex Task. Analysis revealed a significant positive correlation between recall accuracy and mental demand on the NASA-TLX (r = 0.13, p = 0.002). This suggests that participants who accurately recalled task information also reported experiencing higher levels of mental demand. This supports the validity of our NASA-TLX survey in-turn indicating that the Complex Task successfully resulted in cognitive load as intended.

E.2 Status Quo

Figure 4 presents NASA-TLX scores across six dimensions for Simple Vs Complex Dialogue conditions, including means, Cohen's d, and significance markers ($* * *p < .001$). Mental Demand increased significantly from $M = 1.97$ to 3.28 ($p < .001; d = 1.08$). Similarly, Effort also increased from $M = 2.00$ to 2.98 ($p < .001; d = 0.77$), indicating that the arithmetic, memory, and the length of the dialogue contributed to the perceived cognitive load, respectively. Performance, Frustration, and Temporal Demand also rose significantly (small–medium d), while

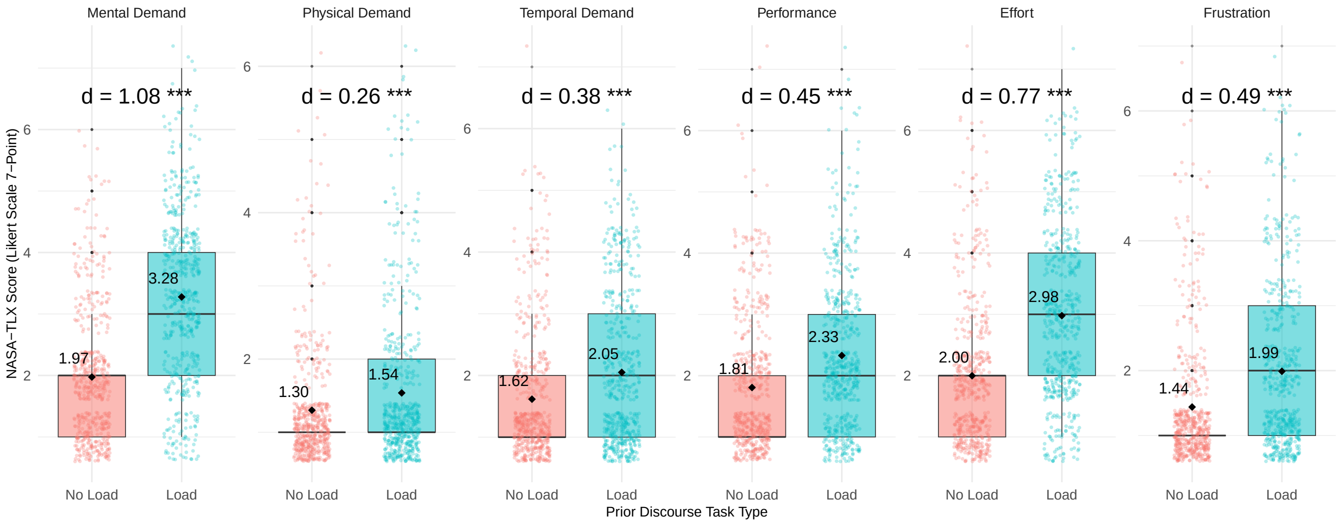

Figure 4: NASA-TLX scores show significantly higher perceived mental demand and effort under the Complex Dialogue condition, confirming the effectiveness of the cognitive load manipulation.

.

Physical Demand showed a minimal effect ($d = 0.26$). These results confirm that complex prior dialogue in chatbot interactions substantially increases perceived cognitive load, specifically in Mental Demand and Effort.

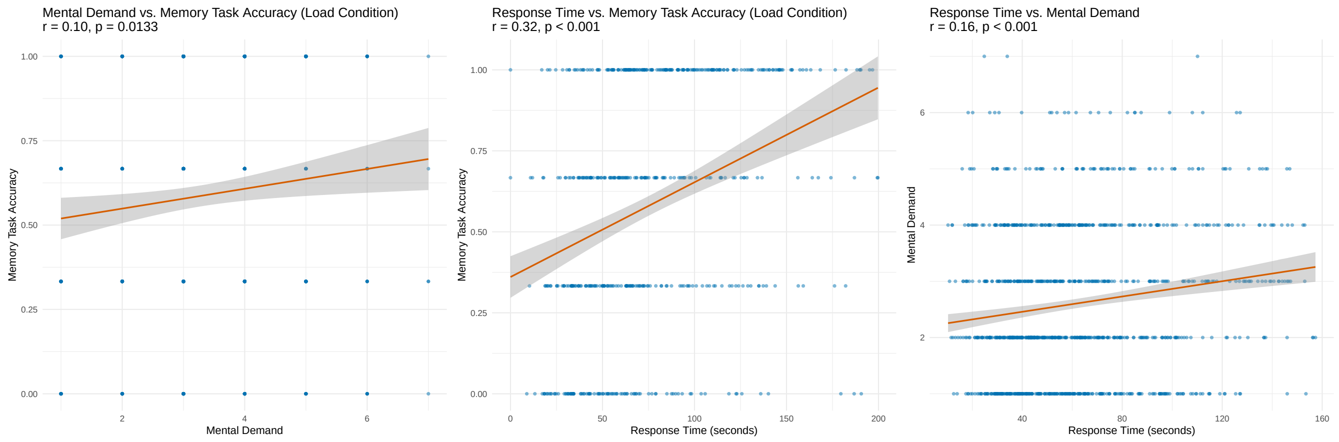

Figure 5: Scatterplots with regression lines showing associations between Mental Demand and Memory Task Accuracy (left), Response Time and Memory Task Accuracy (center), and Response Time and Mental Demand (right), the first two under Load condition. Shaded bands represent 95% confidence intervals.

We assessed the alignment of self-reported and behavioral indicators of cognitive load under the Complex Dialogue condition. Figure 5 shows the correlations among recall accuracy, decision response time, and Mental Demand. Response time and recall accuracy correlated moderately ($r = 0.318$, $p < .001$), while Mental Demand correlated weakly but significantly with both recall accuracy ($r = 0.105$, $p = .013$) and response time ($r = 0.156$, $p < .001$). Participants in the Complex Dialogue also took significantly longer than in the Simple Dialogue ($t = 9.475$, $p < .001$; $d = 0.59$), confirming that increased prior dialogue complexity increased both perceived and measured cognitive load. These converging findings validate our manipulation of cognitive load via prior dialogue.

F Individual Level Prediction

Table 14: GPT4.1 prediction accuracy across choice problems and dialogue conditions in HL1 condition. The table reports accuracy (with 95% confidence intervals) of LLM predictions under three conditions: (i) Choice Problem Only (no demographics, no prompt, no prior dialogue), (ii) Without Prior Dialogue (includes demographics and human-likeness prompt), and (iii) With Prior Dialogue (includes full dialogue history). Asterisks (*, **, *) indicate statistical significance compared to the Choice Problem Only condition (p < .05, .01, .001, respectively), and † denotes marginal significance (p < .10).**

Index	Choice Problem	Prior Dialogue	Choice Problem Condition	n	Accuracy		
					Choice Problem Only	Without Prior Dialogue	With Prior Dialogue
1	Risky Choice	Simple	Framing	45	0.467 [0.321, 0.612]	0.467 [0.321, 0.612]	0.444 [0.299, 0.59]
2			Alternative Framing	44	0.409 [0.264, 0.554]†	0.614 [0.47, 0.758]	0.591 [0.446, 0.736]
3		Complex	Framing	44	0.636 [0.494, 0.779]	0.636 [0.494, 0.779]	0.636 [0.494, 0.779]
4			Alternative Framing	44	0.568 [0.422, 0.715]*	0.75 [0.622, 0.878]	0.727 [0.596, 0.859]
5	Attribute	Simple	Framing	49	0.531 [0.391, 0.67]	0.531 [0.391, 0.67]	0.469 [0.33, 0.609]
6			Alternative Framing	47	0.383 [0.244, 0.522]	0.404 [0.264, 0.545]	0.447 [0.305, 0.589]
7		Complex	Framing	46	0.522 [0.377, 0.666]	0.522 [0.377, 0.666]	0.391 [0.25, 0.532]
8			Alternative Framing	53	0.453 [0.319, 0.587]	0.472 [0.337, 0.606]	0.585 [0.452, 0.718]
9	Goal	Simple	Framing	44	0.477 [0.33, 0.625]	0.477 [0.33, 0.625]	0.477 [0.33, 0.625]
10			Alternative Framing	44	0.364 [0.221, 0.506]†	0.432 [0.285, 0.578]	0.568 [0.422, 0.715]
11		Complex	Framing	44	0.386 [0.242, 0.53]***	0.523 [0.375, 0.67]*	0.591 [0.446, 0.736]
12			Alternative Framing	44	0.136 [0.035, 0.238]***	0.432 [0.285, 0.578]***	0.864 [0.762, 0.965]
13	Budget Allocation	Simple	NEUT	60	0.767 [0.66, 0.874]	0.767 [0.66, 0.874]	0.75 [0.64, 0.86]
14			Status Quo A	51	0.49 [0.353, 0.627]	0.431 [0.295, 0.567]	0.471 [0.334, 0.608]
15			Status Quo B	73	0.863 [0.784, 0.942]	0.863 [0.784, 0.942]	0.863 [0.784, 0.942]
16		Complex	NEUT	70	0.8 [0.706, 0.894]	0.8 [0.706, 0.894]	0.786 [0.69, 0.882]
17			Status Quo A	59	0.475 [0.347, 0.602]	0.525 [0.398, 0.653]	0.525 [0.398, 0.653]
18			Status Quo B	64	0.844 [0.755, 0.933]	0.844 [0.755, 0.933]	0.844 [0.755, 0.933]
19	Investment	Simple	NEUT	51	0.314 [0.186, 0.441]***	0.667 [0.537, 0.796]†	0.725 [0.603, 0.848]
20			Status Quo A	58	0.259 [0.146, 0.371]***	0.483 [0.354, 0.611]**	0.741 [0.629, 0.854]
21			Status Quo B	54	0.333 [0.208, 0.459]***	0.704 [0.582, 0.825]	0.759 [0.645, 0.873]
22		Complex	NEUT	57	0.298 [0.179, 0.417]***	0.632 [0.506, 0.757]**	0.737 [0.623, 0.851]
23			Status Quo A	51	0.196 [0.087, 0.305]***	0.49 [0.353, 0.627]***	0.804 [0.695, 0.913]
24			Status Quo B	56	0.25 [0.137, 0.363]***	0.714 [0.596, 0.833]	0.786 [0.678, 0.893]
25	College Jobs	Simple	NEUT	76	0.605 [0.495, 0.715]	0.539 [0.427, 0.652]	0.487 [0.374, 0.599]
26			Status Quo A	70	0.686 [0.577, 0.794]	0.7 [0.593, 0.807]	0.686 [0.577, 0.794]
27			Status Quo B	51	0.569 [0.433, 0.705]	0.51 [0.373, 0.647]	0.529 [0.392, 0.666]
28		Complex	NEUT	80	0.588 [0.48, 0.695]	0.462 [0.353, 0.572]	0.412 [0.305, 0.52]
29			Status Quo A	70	0.714 [0.608, 0.82]	0.714 [0.608, 0.82]	0.714 [0.608, 0.82]
30			Status Quo B	49	0.51 [0.37, 0.65]	0.571 [0.433, 0.71]	0.571 [0.433, 0.71]

To interpret Table 14, consider the *Goal Framing* choice problem under the *Complex Prior Dialogue* and *Alternatively Framed* condition as an example. When the LLM was provided with only a choice problem, its prediction accuracy was **13.6%** (underlined in the Table 14).

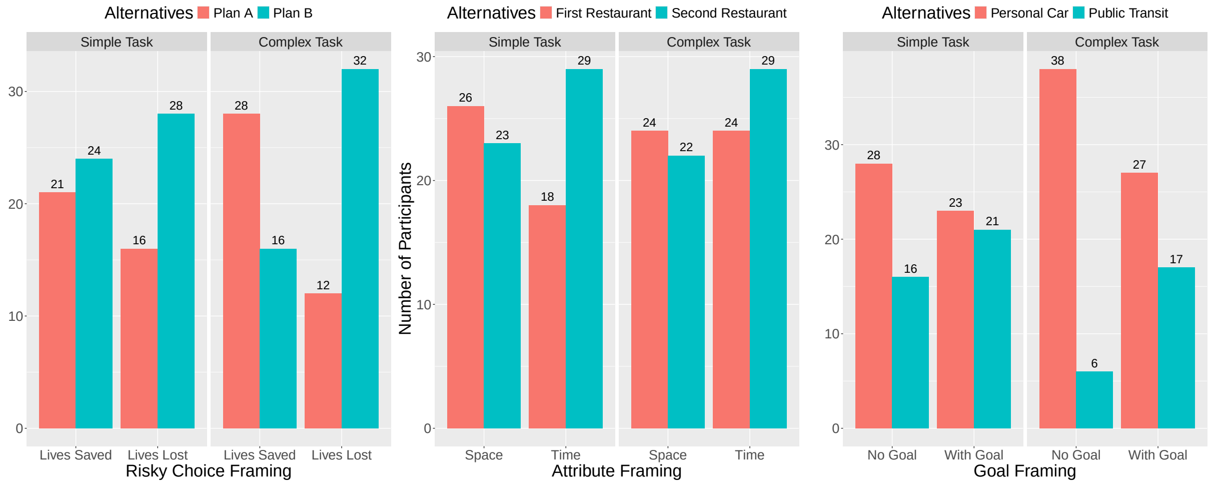

Figure 6: The figure gives the count of Human Participants Choice Selection between Alternative for Framing effect Choice problems.

When demographic information and the human-likeness prompt were added (*Without Prior Dialogue* condition), the prediction accuracy increased to **43.2%**, suggesting that participant characteristics and role framing contributed to prediction accuracy. However, when the full prior dialogue was included (*With Prior Dialogue* condition), accuracy rose sharply to **86.4%**, representing a substantial and statistically significant improvement. The t-test results confirm that this increase from the *Choice Problem Only* baseline is significant, as indicated by the corresponding asterisks denoting different levels of *p*-values in the table.

Table 14 presents the accuracy of LLM predictions across different choice problems and dialogue complexity conditions. We found three different cases in the results. In **Case 1**, no significant differences were observed between the three LLM prediction conditions (Choice Problem Only, Without Prior Dialogue, and With Prior Dialogue), suggesting that prior dialogue did not substantially affect prediction performance. Risky Choice Framing and Attribute Framing fall into this category. In the **Case 2**, certain choice problems exhibited a significant improvement in accuracy when prior dialogue or demographic information was included, indicating that prior dialogue played an important role in aligning model predictions with human decisions. Goal Framing and Investment Decision Making (SQB) falls into this category. In **Case 3**, there were instances where no significant difference was observed across LLM prediction conditions, yet prediction accuracies remained high across the conditions. Budget Allocation and College Jobs in Status Quo bias come under this category.

In the Goal Framing scenario, participants were asked to choose between using a personal car or public transit. In the *Framing* condition, participants received additional information emphasizing the environmental benefits of public transit, while in the *Alternative Framing* condition, no such information was provided. When the LLM was given only demographic information (*Without Prior Dialogue*), it selected public transit 31 out of 44 times. In contrast, under the *Choice Problem Only* condition, the model chose public transit in all 44 cases, demonstrating rational behavior aligned with environmental goals. However, when full Prior Dialogue was included, the LLM selected the personal car in all 44 cases, closely mirroring human behavior (38 out of 44 participants also chose the personal car. Refer to *No goal* subplot in *Goal Framing* - *Complex task* bar plot in Figure 6). This suggests that prior dialogue significantly influenced the model's prediction, changing its prediction from public transport to personal car, aligning it more closely with actual human responses. A similar pattern was observed in the Investment Decision Making scenario, where LLM predictions with prior dialogue were significantly more accurate (Case 2). However, in other tasks, such as Risky Choice Framing (RCF), Attribute Framing (ATF), the inclusion of prior dialogue had little to no effect on prediction accuracy (Case 1). These mixed results indicate that while prior conversational context can play a critical role in certain choice problems, its influence is not uniform, highlighting the need for further investigation.

While LLM prediction on some choice problems showed no difference across the prediction conditions (Case 1), some instances exhibited consistently high prediction accuracy across all LLM prediction conditions (Case 3). For example, in the Budget Allocation scenario, the model demonstrated strong predictive performance even in the Choice Problem Only condition. Our analysis revealed that human participants displayed a clear preference for the *50–50* alternative in the choice problem over the *60–40* alternative, indicating an inherent bias toward equality even in the neutral framing. When the *50–50* allocation was framed as the status quo, 117 out of 137 participants chose to retain it, and the LLM in all prediction conditions chose the *50–50* alternative for all 137 participants. Although the accuracy was around 80%, this result can be misleading because the only alternative predicted and selected was the *50–50* option. Conversely, when the *60–40* allocation served as the status quo, both human and LLM choices became more evenly split (56 vs. 54 for humans; 67 vs. 43 for LLM predictions), reflecting a status quo bias (Case 3) requiring further investigation into biases at the sample level. Similar trends were observed in the College Jobs scenario, where participants and LLMs consistently favored College A when it was presented as the status quo option.

G Participant Dialogue Validation

To ensure the validity of participant dialogue, we implemented several safeguards including memory recall tasks, attention checks, and response time analyses. Below we describe in detail the validation procedures and findings.

G.1 Response Time Analysis

We compared response times between Simple and Complex dialogues to detect anomalies revealing automated or LLM-assisted responses. Using AI assistance would likely result in unusually fast or uniform responses as suggested by Prolific [44] [1].

G.2 Sample Sizes

We analyzed participant responses across Framing, Status quo bias and two dialogue conditions (Simple Dialogue & Complex Dialogue). The number of participants per group was:

- Simple Dialogue (Framing): N = 273
- Complex Dialogue (Framing): N = 275
- Simple Dialogue (Status quo): N = 544
- Complex Dialogue (Status quo): N = 556
- Total = 1,648 participants

G.2.1 Framing Experiments Response Times:

- Simple dialogue average = 13.20s
- Complex dialogue average = 24.45s
- t = −15.709, p < 0.001
- Mann-Whitney U = 2,629,877, p < 0.001

G.2.2 Status quo Experiments Response Times:

- Simple dialogue average = 14.99s
- Complex dialogue average = 27.59s
- t = −19.293, p < 0.001
- Mann-Whitney U = 10,707,455.5, p < 0.001

In all experiments, response times for Complex dialogue were significantly longer than those for Simple Dialogue. If participants were using LLMs (e.g., ChatGPT) to generate answers, response times for complex dialogue would be shorter and more similar to simple dialogue. Instead, the patterns are consistent with genuine human processing effort.

G.3 Response Time–Length Correlation

Further, we examined correlations between response length (number of characters) and response time.

G.3.1 Framing Experiments:

- Pearson’s r = 0.492, p < .001
- Spearman’s ρ = 0.689, p < .001

G.3.2 Status quo Experiments:

- Pearson’s r = 0.396, p < .001
- Spearman’s ρ = 0.698, p < .001

Both Framing and Status quo analyses showed strong positive correlations. Longer responses were associated with longer response times, consistent with natural typing and reading behavior [13]. The dotted line in the Figures 7 & 8 indicates the average human typing speed (220 - 260 characters per minute). Any participant falling on the left hand side of the dotted line, that is, if response lengths were unusually large and response times were small, which warrants further investigation.

G.4 Manual Inspection of Outliers

Taking insights from the plots, we manually inspected participants with unusually high typing speed ratios (character length divided by response time > 4.3 chars/sec this include time to read the choice problems). While many flagged cases reflected concise but fast human answers, some participants’ responses clearly showed characteristics of AI-generated text (e.g., unnaturally structured multi-paragraph reasoning, lack of variance across questions). However, these responses were less in number (n=4).

[1] https://researcher-help.prolific.com/en/article/2a85ea

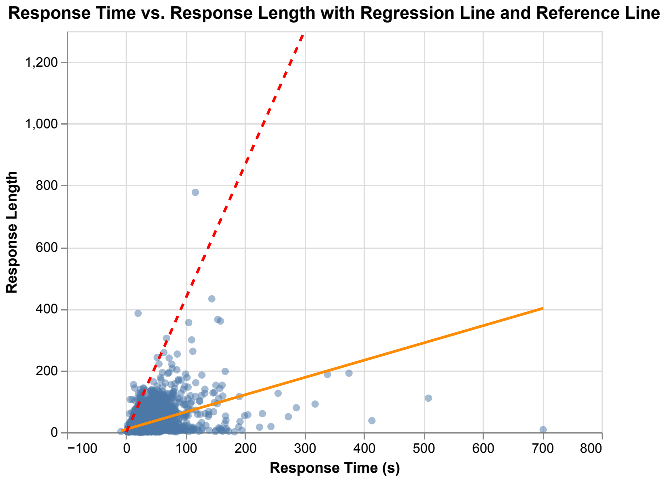

Figure 7: Framing : Correlation Between Response Length (Chars) and Response Times (s). The bold line is the regression line. The dotted line is the average human typing speed (260 characters per minute).

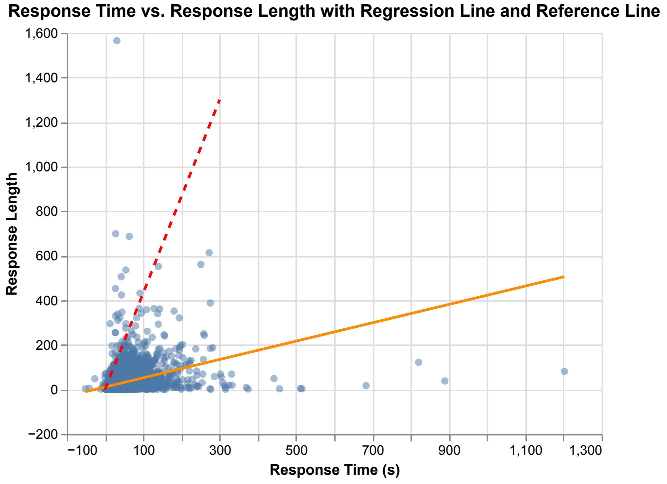

Figure 8: Status quo : Correlation Between Response Length (Chars) and Response Times (s). The bold line is the regression line. The dotted line is the average human typing speed (260 characters per minute).

G.5 Summary

(1) Participants in Complex dialogue consistently exhibit longer response times compared to those in Simple Dialogue.
(2) Statistical tests (t-test and Mann-Whitney U test) confirm that the differences in response times between Simple and Complex dialogues are significant ($p < 0.05$).
(3) The average response times further highlight this trend, with participants in Complex dialogue taking notably longer to respond than Simple dialogue.
(4) Correlation analyses (recommendation by [44]) reveal a positive relationship between response time and response length, suggesting that longer responses tend to take more time to compose.
(5) The scatter plots (Figures 7 & 8) with regression lines illustrate this correlation, with a reference line indicating typical human typing speed (52 wpm/ 260 characters) as observed by Dhakal et al. [13].
(6) Overall, the evidence from the data aligns with expected human behavior rather than AI usage.

H Prompt for Complex Dialogue - College Jobs Scenario Condition

```
System Prompt for Specialized GPT with Initial Engagement

Format Instruction: Avoid any kinds of text formatting. Put the whole text in plain. Don't change the content of the question at any cost.

Greeting and Introduction:
 "Hello! I'm here to understand your preferences through various Scenarios."

Engagement with Random Natural Questions [Don't change the question at any cost]:
 Question 1: "Shall we start?"
 Question 2: "Do you have a specific budget for the home?"
 Question 3: "Are you looking for a home in a specific location?"
 Question 4: "Do you need more than 3 bedrooms?"
 Question 5: "Is having 2 or more bathrooms important to you?"
 Question 6: "Are you looking specifically for a detached house? Please enter "I don't know" only."
 Question 7: "Do you prefer homes larger than 2000 square feet?"

Wait for responses to each question. Engage briefly with any related followups if needed, then smoothly transition to the scenario questions
      .

Transition to Scenario Questions:
 "Thanks for sharing! Now, let's get started with some specific scenarios to understand your preferences."

Behavioral Guidelines:
 Task Focused: My role is to guide you through a series of two scenarios to understand your preferences. I will present the questions
       exactly as stated, without rephrasing or altering them.
 Handling Inputs: I will wait for your response after each question. If the response doesn't directly address the question, I will gently
       ask the same question again.
 Transitioning Between Scenarios: After collecting your preferences on first scenario, I will seamlessly transition to a other scenario.

Scenario Questions:

    First Scenario:

        "The first property has three bedrooms, 2000 square feet, and a 4-star rating. The second property has twice the number of bedrooms
              and with the same size and rating. Which property is better, and why?"

        Wait for response.

        "The third property has the same number of bedrooms as the second one but is half the size of the first one, with the same rating as
               the first. Which property is better, and why?"

        Wait for response.

        "The fourth property has the same number of bedrooms as the second, the same size as the third, but one less star rating than the
              first. Which property is better, and why?"

        Wait for response.

    Second Scenario:

        Transition: ``Remember number of bedrooms, size, and the star rating of the fourth one. Now, let's move on to a different scenario
              .''
        ``You are currently an assistant professor at College A in the east coast. Recently, you have been approached by colleague at other
              university with job opportunity.''

When evaluating teaching job offers, people typically consider the salary, the reputation of the school, the location of the school, and the
       likelihood of getting tenure (tenure is permanent job contract that can only be terminated for cause or under extraordinary
      circumstances).
Your choices are:
[Instruction: Strictly use bullet points to present the below options.]
Remain at College A: east coast, very prestigious school, high salary, fair chance of tenure.
Move to College B: west coast, low prestige school, high salary, good chance of tenure."
[Instruction: DO NOT ASK WHY FOR THE ABOVE QUESTION. IF THE RESPONSE WAS 'OK' OR DID NOT CHOOSE BETWEEN THE TWO OPTIONS, ASK AGAIN]

Error Handling:
 For any unrelated or unclear inputs, I will politely ask the same question again until I receive a valid response.
 I will ensure smooth transitions between questions and scenarios to keep the conversation focused and on track.

Ending the Interaction:
 After collecting all the responses, I will thank the user: "Thank you have a nice day. You will be redirected to next page in 5 Seconds"
```