# AI Agents Can Enable Superior Market Designs*

Gili Rusak† Harvard
Benjamin S. Manning† MIT
John J. Horton MIT & NBER

March 19, 2025

**Abstract**

Many theoretically appealing market designs are under-utilized because they demand preference data that humans find costly to provide. This paper demonstrates how large language models (LLMs) can effectively elicit such data from natural language descriptions. In our experiment, human subjects provide free-text descriptions of their tastes over potential roles they could be assigned. An LLM converts these descriptions into cardinal utilities that capture participants' preferences. We use these utilities and participants' stated preferences to facilitate three allocation mechanisms—random serial dictatorship, Hylland-Zeckhauser, and a conventional job application type game. A follow-up experiment confirms that participants themselves prefer LLM-generated matches over simpler alternatives under high congestion. These findings suggest that LLM-proxied preference elicitation can enable superior market designs where they would otherwise be impractical to implement.

*Manning thanks Tyler Cowen and the Mercatus Center for their generous funding and intellectual support. Horton thanks Upwork for their generous funding and support. All authors thank David Parkes for his helpful comments and suggestions. Author contact information, code, and data are currently or will be available at `http://www.benjaminmanning.io/`. Horton is a co-founder of a company, Expected Parrot Inc., using generative AI models for research. Manning also holds a financial interest in Expected Parrot Inc.

†Both authors contributed equally to this work—the order of their names is random.

# 1 Introduction

Many theoretically appealing market designs remain surprisingly underutilized in high-stakes settings. A key obstacle to their broader adoption is that these mechanisms often impose extensive informational and cognitive demands on participants, who must articulate precise preferences over dozens, hundreds, or even thousands of items. Humans, however, do not usually have the time nor patience to make such fine-grained comparisons at scale, and even if they do, they are prone to mistakes (Enke, 2024; Oprea, 2024). While small errors in preference reporting might seem innocuous by the logic of the envelope theorem, even minor deviations can degrade equilibrium outcomes (Akerlof and Yellen, 1985), undermining the hypothetical benefits of these designs.

If burdensome preference elicitation is indeed the bottleneck to implementing superior mechanisms (e.g., Budish (2011)), recent advances in machine learning suggest a potential remedy (Brero et al., 2021). In particular, large language models (LLMs) can be prompted to respond to human desires described in natural language (Bubeck et al., 2023). If such descriptions capture key aspects of human *tastes*—the fundamental features of utility functions that shape preferences (Stigler and Becker, 1977; Cowen, 1989)—then LLMs could be powerful tools for evaluating large menus of options. Operating at effectively zero marginal cost, they could reduce errors and save valuable human time.

For example, imagine someone describing their fruit tastes as: *"I love crisp textures and tart flavors but avoid fruits with pits."* Such a statement is effectively a flexible program for ranking all sorts of fruit, not just a single ordinal checklist. Given a set of fruit, these tastes imply a natural ordering (for instance, apple $\succ$ banana $\succ$ cherry), and an LLM given this description could assign cardinal utilities (e.g., $u(\text{apple}) = 1$, $u(\text{banana}) = .4$, $u(\text{cherry}) = 0$), preferences over bundles, and so on. Moreover, if a new fruit, say, a date, is introduced, the LLM can infer that it has a pit but is not tart like some cherries, reporting the user's preference appropriately (cherry $\succ$ date).[1]

Preliminary work suggests that LLMs can indeed convert textual tastes into precise pref-

[1] Figure A2 shows that GPT-4o correctly orders all four fruits given this exact taste description.

erence statements (Li et al., 2023). Yet two critical research questions remain: (i) whether individuals can effectively convey their complex tastes in a short paragraph, and (ii) whether an LLM's conversion of those tastes into preferences is accurate enough to improve real economic outcomes.

To examine these questions, we use a labor-market-like matching experiment involving crowdsourcing workers (Horton et al., 2011) recruited via Prolific. We focus on labor markets for several reasons. First, it provides a realistic, high-stakes environment in which participants have strong, varied preferences and face challenging matching constraints. Furthermore, when too many job seekers apply to the same positions—a pervasive problem in modern labor markets (Horton et al., 2024; Fradkin et al., 2024)—participants must strategically decide where to focus their ranking effort, considering not only their own preferences but also likely competition from others. This makes accurate preference elicitation both more valuable and more challenging. Finally, labor markets are often dynamic, with new jobs appearing and old ones disappearing, making static rankings of old options quickly obsolete. By contrast, an LLM endowed with a participant's written taste description can flexibly compute preferences for new jobs without needing a fresh ordinal list.

In our experiment, each of $n = 781$ participants wrote a short natural language description of the Prolific tasks they enjoy and then rank-ordered 50 representative tasks. Our goal is to assign participants tasks subject to a labor market-esque one-to-one matching constraint with a variety of allocation mechanisms. We translated these descriptions into von Neumann-Morgenstern (vNM) utilities by prompting an LLM—endowed with the participant's stated tastes—to evaluate each task under a structured lottery (von Neumann and Morgenstern, 1944).

Evaluating the LLM's performance, however, is not straightforward. Participants' rankings do not provide an unambiguous source of ground truth. Our tracking software reveals a clear pattern: participants tend to focus on ordering their top choices while often leaving most other tasks in their initial random positions. Considering that task order is initially random for each participant, this suggests that participants are not reporting their full preferences.

In this sense, we as researchers are facing the very problem we hope to solve: high-quality preference data are needed to validate the LLM, yet humans themselves provide incomplete rankings.

To be clear, the reported rankings provide non-trivial information about participants' preferences, and the rankings implied by the LLM-derived utilities correlate strongly with these rankings. Furthermore, the more effort participants put into ranking tasks, and therefore, providing a better measure of ground truth, the better the LLM's performance. If anything, this suggests that human preferences likely understate the LLM's performance.[2]

To address the evaluation problem posed by under-sorted rankings, we implemented a second phase of the experiment in which the same participants from the initial ranking exercise make direct choices from a very limited set of tasks assigned by different mechanisms. Some mechanisms are facilitated by participants' own self-reported rankings, while LLM-derived preferences facilitate others. We can then compare participants' choices between these tasks as a function of the preference data used to generate them. Because participants only choose among a few tasks, cognitive demands are minimal, giving us a clearer, ground-truth measure of how effectively the LLM-based approach aligns with real preferences.

Constructing matches for the follow-up experiment requires several steps. Guided by participants' (imperfect) rankings and the LLM-derived utilities, we first create artificial markets of 50 participants each to study how varying degrees of preference overlap affect outcomes. We imagine our Prolific workers in a conventional job market where each worker can only be assigned to one task. Specifically, for each participant, we create two distinct markets: one in which they are grouped with 49 participants whose preference lists heavily coincide (a *congested* scenario), and another with 49 participants whose preferences rarely overlap (an *uncongested* scenario). By varying the degree of preference overlap, we can compare the relative tradeoffs between human and LLM-derived preferences in environments

[2]One might suspect that raising the stakes would induce better ranking. In a companion study, we demonstrate that this is not the case. We used our preference elicitation interface to ask MBA students to rank 31 assignments that would determine their semester-long course projects; under-sorting persisted despite the smaller set of items, days of deliberation, and real academic consequences (see Appendix A). Even strong incentives, a far smaller set of items, and sufficient time do not facilitate fully elicited preferences.

where we might expect one to be more effective than the other.

Within each of the participant's two artificial markets, we then simulate three mechanisms. The first is ordinal-based random serial dictatorship, using each participant's own reported rankings. The second is cardinal-based Hylland-Zeckhauser pseudo-market (Hylland and Zeckhauser, 1979), based on the LLM-derived utilities. The third is a maximizing and more straightforward "conventional job application" game in which each participant picks exactly one task, bearing the risk of being crowded out if many people select the same top choice.

Two days after the initial ranking exercise, participants returned to select from three tasks—one from each mechanism—drawn from either their *congested* or *uncongested* market, depending on random assignment. Indeed, the findings show that when preference overlap is light, participants tend to favor the simpler job application game, which often yields their top choice. Under heavy competition, however, they gravitate toward either the LLM-driven cardinal mechanism or random serial dictatorship, providing direct causal evidence that LLM-derived utilities can effectively capture participants' preferences in a realistic, market-like environment. To reiterate, by allowing participants to choose among just three tasks (blinded to which mechanism generated each match), we confirm that the LLM-based approach can successfully generate matches that participants prefer.

Having established that participants prefer LLM-generated matches under high congestion, we now analyze the welfare implications more systematically using our market simulations. In the uncongested market simulations—where top preferences rarely coincide—simple ordinal mechanisms generally suffice, allowing most participants to secure their first choice. By contrast, in congested markets, the cardinal Hylland-Zeckhauser mechanism achieves substantially higher overall welfare, as measured by the LLM-derived utilities. Incorporating richer vNM utilities—obtained at negligible marginal cost through an LLM—helps mitigate the negative consequences of crowding around the same tasks. Intuitively, if participants' lists rarely overlap, almost everyone secures a well-sorted top pick anyway; but with severe overlap, these more granular preferences become invaluable.

One concern is that these welfare improvements might be an artifact of using the same

LLM-based utilities for both allocation and evaluation. Yet the LLM's strong alignment with participants' top-ranked items suggests that the observed gains are not merely circular. Nonetheless, we cannot guarantee perfect accuracy, which is why the follow-up experiment serves as a crucial check on our results more generally.

Although these findings underscore the potential of LLMs as agents for preference elicitation, there remains considerable room to improve the alignment between an LLM's interpretation of user tastes and users' true preferences. Even with the participant under-sorting, it is clear that the LLMs are not perfect. Indeed, previous work has shown that LLMs can fail in unexpected ways (Vafa et al., 2024). Moreover, scaling these methods to larger populations and more complex matching scenarios introduces significant computational and design challenges (Vazirani and Yannakakis, 2020; Aziz et al., 2013).

Our work builds on several intersecting research areas. First, we extend the emerging literature examining LLMs' capacity to approximate human-like behavior in various economic contexts (Manning et al., 2024; Horton, 2023). A complementary line of research explores AI-assisted evaluations and negotiation (Lira et al., 2023; Tessler et al., 2024).

Our work also contributes to the evolving discourse on alternative preference elicitation methods. Recent studies suggest that qualitative self-assessments may serve as efficient substitutes for traditional preference elicitation techniques (Falk et al., 2018, 2022), offering cognitive advantages similar to our taste descriptions. However, debate persists about whether these assessments measure the same constructs as conventional methods (Chapman et al., 2025). We offer an alternative path forward: rather than proxying traditional preference elicitation with qualitative assessments, we use LLMs to bridge the gap—converting natural language taste descriptions into precise utility functions consistent with traditional economic frameworks.

Most directly related to our approach is the nascent literature on AI-assisted preference elicitation. Previous work has explored AI applications in social choice, auctions, lotteries, and recommendation systems (Huang et al., 2025; Li et al., 2023; Kim et al., 2024; Soumalias et al., 2025; Brero et al., 2021). However, these studies rely predominantly on

simulations or small-sample experiments. Our contribution extends this work in three significant ways. First, we provide robust evidence from a large-scale experiment demonstrating that LLM-derived utilities can effectively represent human preferences in realistic market settings. Second, we show that these automatically derived cardinal preferences enable the implementation of sophisticated allocation mechanisms that outperform simpler alternatives under market congestion. Finally, we offer direct experimental confirmation that participants themselves often prefer matches generated through LLM-derived preferences, particularly in congested environments.

Taken together, our findings suggest that LLM-based preference elicitation can substantially reduce the cognitive and temporal barriers to detailed preference reporting. By converting brief natural language descriptions into comprehensive utility functions, this approach enables the practical deployment of theoretically superior market designs that have previously been limited by the challenges of obtaining granular preference data at scale.

## 2 Experiment design

We design and execute a two-phase study to evaluate whether participants can effectively communicate their tastes through short natural language descriptions and whether LLMs can accurately translate these descriptions into preferences suitable for matching. In Phase I, participants provide a free-text narrative of their task preferences and then rank a set of representative Prolific tasks. Phase II is a randomized experiment that asks the same participants to choose among tasks derived from different allocation mechanisms—some based on their own preferences, others based on the LLM-derived utility—thereby offering a direct comparison of the mechanisms' outcomes from participants' own perspectives.

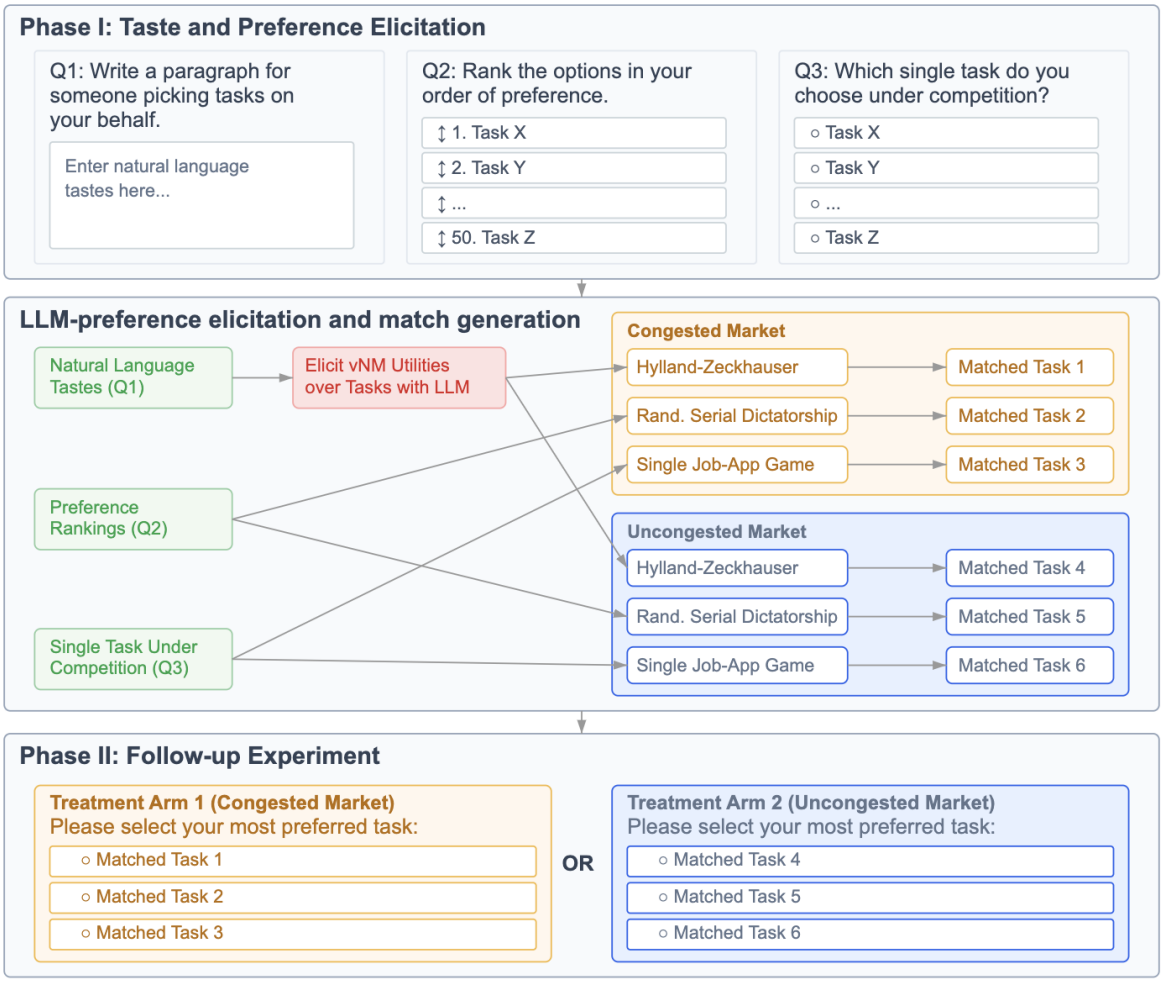
### 2.1 Phase I: taste and preference elicitation

Phase I of the experiment comprised three questions given to $n = 781$ participants using a custom web interface.[3] The top panel of Figure 1 shows the phase I flow. First, participants

[3]We excluded 19 participants due to random errors in data recording from our online interface.

were asked to describe their tastes in tasks in natural language (Q1 in Figure 1). Specifically, they were instructed to provide written guidance to a friend selecting Prolific tasks on their behalf; both features of tasks they liked and disliked were elicited in short paragraphs. Participants were incentivized with a bonus for accurately describing their tastes relative to the rankings they would subsequently submit.

**Figure 1:** Experimental flow




*Notes: This figure illustrates our two-phase experiment. The top panel shows Phase I, where participants (i) describe their tastes in a free-text prompt (Q1), (ii) rank 50 tasks (Q2), and (iii) choose exactly one task as their "single job application" (Q3). The middle panel outlines how these data are used to generate cardinal utilities via the LLM and to simulate three mechanisms—random serial dictatorship, the single job application game, and Hylland-Zeckhauser—within both congested and uncongested markets. The bottom panel depicts Phase II, in which participants return to pick their preferred match from among the resulting tasks in a brief follow-up study.*

Second, participants were presented with an interface that asked them to rank a set of

50 randomly arranged Prolific tasks in order from most preferred to least (Q2 in Figure 1). Users could click on the tasks to see details on time commitments and compensation and could re-rank items as much (or as little) as they desired. Each task was designed to reflect a realistic Prolific task. They varied along three dimensions: hourly rates ranging from $10 to $20, durations from 5 to 45 minutes, and different types of expectations for the work required. For example, some tasks offered participants the opportunity to solve puzzles, others to reflect on daily activities, and others to share opinions on politics. The full set of tasks is provided in the appendix along with screenshots of the ranking interface.[4]

Finally, participants were taken to a new page (Q3 in Figure 1) and asked to select exactly one of the 50 tasks, under the assumption that other participants were facing the same decision problem. The instructions explained that if a participant were the only person to choose a particular task, they would receive it automatically. However, if multiple participants chose the same task, one would be randomly selected to receive it, and the rest got no assignment. This *single job application* game is analogous to a real-world scenario in which an individual can apply for only one position while facing potential competition from other candidates. We refer to the choice made by participants in this step as their "single job application" herein.

## 2.2 LLM-preference elicitation and match generation

The middle panel of Figure 1 outlines how we process the Phase I data to generate task matches for the participants. We begin by transforming each participant's natural language descriptions (Q1) into numerical vNM utilities. Specifically, we fed each description to an LLM and prompted it to score how much a participant would value each of the 50 tasks, creating a cardinal preference profile for each individual. This new set of utilities effectively serves as a third preference measure alongside the ordinal rankings (Q2) and the single job

[4]Two example tasks include: (i) Navigation Task—Plan routes and make directional decisions in a map-based exercise designed to assess your spatial awareness. This should take about 35 minutes and pays 7.50 dollars. (ii) Digital Literacy Survey—Share your experiences with various digital tools and platforms through a survey that examines your overall comfort with technology. This should take about 20 minutes and pays 4.00 dollars.

application (Q3). We describe this entire process, along with the other matching details below fully in Section 4.

Next, we construct two distinct market scenarios for every participant. In a *congested* market, they are grouped with participants whose Q2 lists overlap heavily, creating an environment where many participants compete for the same top-ranked tasks. In an *uncongested* market, they are grouped with participants whose rankings differ substantially, minimizing the risk that many people chase the same task.

Finally, we simulate three different mechanisms in both markets. We run random serial dictatorship using each participant's full ordinal list. We also implement a single job application game based on each participant's single job application and use the LLM-derived utilities to allocate tasks via Hylland-Zeckhauser. This yields up to six possible matches per participant: three from their congested market and three from their uncongested market. As shown in the middle and bottom panels of Figure 1, these simulated matches lay the groundwork for the Phase II follow-up, where participants evaluate the assignments themselves.

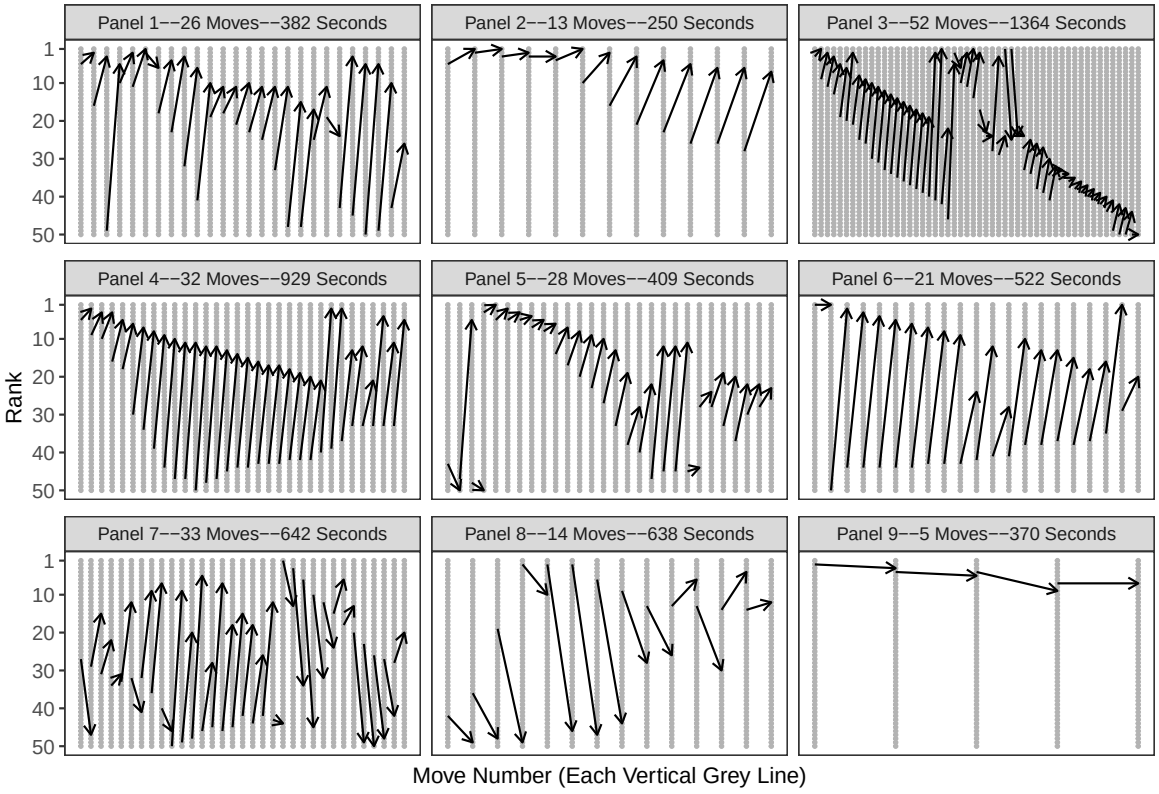
### 2.3 Phase II: follow-up experiment

Two days after Phase I, we invited participants back for a brief follow-up experiment (Figure 1, bottom panel). Of the original 781participants, 704returned to complete this second phase. Each participant was randomly assigned to either a congested or an uncongested treatment condition. They then saw the set of three tasks generated by random serial dictatorship, Hylland-Zeckhauser, and the single job application game under their assigned market type (congested or uncongested) from the matching and preference elicitation simulations. Participants were asked to pick exactly one task they most preferred from those shown. These tasks appeared in random order, with no indication of the relevant mechanism. In some cases, the game did not yield an assignment, or two mechanisms yielded the same task, leaving only two options.

# 3 Phase I results and LLM-preference elicitation

In Phase I, we collected two types of preference data from participants: natural language descriptions of their tastes and explicit rankings of 50 representative Prolific tasks. This section analyzes these data in three parts. First, we document participants' ranking behavior, showing that they significantly under-sort their lists, primarily focusing on ordering only their most preferred tasks. Second, we examine participants' natural language descriptions, which reveal rich and idiosyncratic task preferences. Finally, we detail how we transform these free-text descriptions into cardinal preferences using LLM-elicited vNM utilities and present evidence on the accuracy of these derived preferences.

**Figure 2:** Examples of participant ranking behavior




*Notes: Each panel shows a participant's ranking behavior through directional arrows indicating task repositioning, with task duration and the number of moves made displayed at the top. Each arrow represents a single move, with the base of the arrow indicating the original rank of the moved task and the tip of the arrow indicating the task's new rank.*

### 3.1 Ranking behavior

Figure 2 offers a representative visual snapshot of how participants sorted the tasks according to their preferences. Each panel captures a single participant's sequence of moves over time. Each vertical row of 50 dots represents the 50 tasks in a participant's rankings, with arrows showing how particular tasks were repositioned with each move. For example, in Panel 1, the participant relocated 26 different tasks over 382 seconds, starting by moving the task initially at rank 5 to rank 2, then the task at rank 16 to rank 3, then the task at rank 49 to rank 5, and so on. In this way, the figure highlights the variety of strategies. While some participants made only a handful of moves (e.g., Panel 9), others employed extensive back-and-forth adjustments across the entire list (e.g., Panels 3 and 7).

Despite the diversity in Figure 2, one trend is common across many participants: the top six panels show a clear pattern tracking the heads of the arrows (i.e., the destination rank of the task moved) from the upper left sequentially down to the lower right. This pattern suggests that participants tended to start by moving their most preferred tasks to the top of the list and then move tasks to successively lower ranks (higher numerical ranks). To make this concrete, the simplest version of this strategy would be to move the most preferred task to rank 1, then the second most preferred task to rank 2, and so on, sequentially lower in the rankings. Note that when we say a task is ranked "lower," we mean it has a higher numerical rank (e.g., rank 10 is lower than rank 1). When a task is ranked "higher," the opposite is true.

Of course, participants did not perfectly follow this strategy, but the trend is clear. Further evidence is provided by a simple regression of destination rank on move order, clustering on and controlling for participant fixed effects. The coefficient is positive and precise ($\beta = 0.264$ (0.026)). On average, participants moved tasks to a rank 0.264 slots lower than the rank of the task they moved directly preceding that move. Table A1 in the appendix shows this specification in detail. If participants were merely sorting their lists with maximal efficiency, we would not expect to see such a strong statistical relationship.

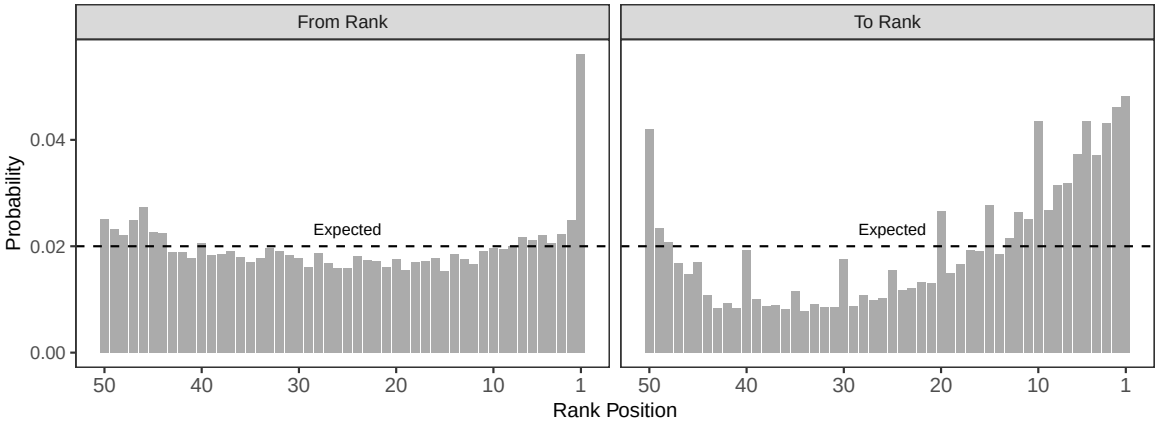
We refer to this heuristic as a "cascading downward" strategy herein. Further evidence

of the pattern is provided in Figure A7 in the Appendix. For example, across all sequences of fives consecutive moves made by all participants, more than 40% of them were monotonically decreasing in rank from the first move to the last.

### 3.1.1 Sorting behavior focused on top positions

To further explore this cascading downward strategy, we separated each reranking action for every participant into two steps: moving a task *from* a rank and moving a task *to* a rank. The left panel of Figure 3 shows that the empirical distribution of movements *from* each rank is close to uniform, This is to be expected if participants were simply sorting their lists with rationally—no rank should be systematically more or less likely to be moved *from*. The large uptick at rank 1 likely reflects participants testing the interface by immediately moving the first most visually salient task at the top of the list.

**Figure 3:** Distribution of how often participants moved tasks from and to each rank position




*Notes: The left panel shows the fraction of task movements originating from each rank, and the right panel shows the fraction of moves arriving at each rank. A perfectly rational participant would move tasks to and from each rank with uniform probability of* $\frac{1}{50} = 0.02$.

In contrast, the distribution of movements *to* each rank in the right panel of Figure 3 exhibits a pronounced checkmark shape. An inordinately large number of moves push tasks to the top 10 positions and a smaller but noticeable fraction lands at the very bottom. Meanwhile, the items in rank 20-45 receive significantly fewer repositionings. This pattern adds further evidence that participants were highly inclined to locate their favorite tasks (and in some cases, identify the handful they liked least) while disproportionately leaving the bulk

of tasks in the middle to lower tasks near their random starting position

### 3.1.2 Participants showed substantial under-sorting

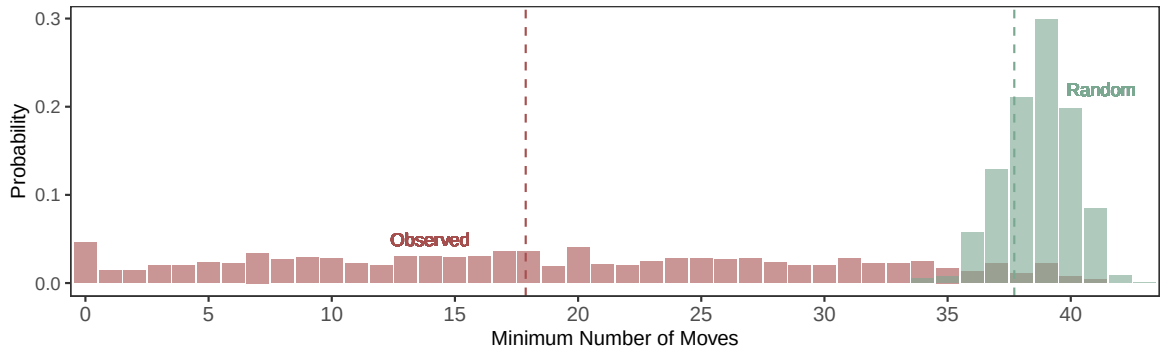
Technically, participants could have disproportionately moved tasks to the top positions but still revealed their true preferences. This would simply mean participants made many redundant moves to the top positions. To provide definitive evidence that participants were under-sorting, we compared their actual reranking behavior to what we would expect from purely random permutations. For each participant, we first calculated the minimal number of insertion moves (Schensted, 1961) necessary to transform the initial (random) list into their final declared submitted list of rankings. This measure reflects a theoretical lower bound on how many operations one should need to position every task correctly, assuming that the final list is fully sorted.

Next, we randomly selected a list of 50 items and then generated 1,000 random permutations of this list. For each permutation, we calculated the minimal insertion moves required to restore the original ordering, creating an empirical distribution of these minimal moves. This distribution serves as our benchmark for how many moves in expectation it takes to fully sort a list of 50 items starting from a random ordering. Assuming that participants have fully specified preferences that they explicitly ranked, we would expect the empirical distribution of minimal moves to converge to this random benchmark even if they were making some redundant moves to the top positions. If redundant moves lead to a fully specified list, they do not affect the distribution of minimal moves since the minimal moves are calculated relative to the original ordering.

Figure 4 displays the empirical distribution of these minimal moves (in red), alongside the simulated benchmark from random permutations (in green). The dashed vertical lines show the average number of moves for each distribution, colored accordingly. In random permutations of 50 items, typically 34-42 moves are needed to reach a fully sorted order with an average of 39. In contrast, participants in our study report final lists requiring far fewer moves to transform from the original lists they were presented. The average number

of minimal moves was around 19, with roughly 90% of participants needing fewer than 34 moves: the lowest number of minimal moves needed to sort the list in the 1000 random permutations. This is unequivocal evidence of severe under-sorting.

**Figure 4:** Distribution of minimal insertion moves to achieve participants' final rankings vs. random permutations




*Notes: The red distribution shows the actual minimal insertion moves needed to turn each participant's randomly ordered tasks into their final ranking, while the green distribution shows the same statistic for purely random permutations of 50 items. The dashed lines show the average number of moves for each distribution.*

Taken together, our analysis of participants' ranking behavior identifies where we should expect participants to have both accurately specified preferences and under-sorted. On the one hand, they seem to have accurately identified and ordered the items they were most enthusiastic about, as reflected in their concentrated moves to the highest ranks. On the other hand, the middle and lower portions of the lists saw far less movement, showing that in the aggregate, participants did not fully reveal their preferences. While such cascading top-down sorting strategies are understandable and are likely effective heuristics for sorting a small set of items, such sorting rules are not scalable.

Importantly, use of this heuristic is not a monolith (although under-sorting in general is). Some participants did not follow this strategy by completely it, cascading up from the bottom ranks (e.g., Panel 8 in Figure 2) rather than down from the top. Others exhibited seemingly random movements with no clear pattern (see Panel 7 in Figure 2). However, both of these behaviors were comparatively rare.

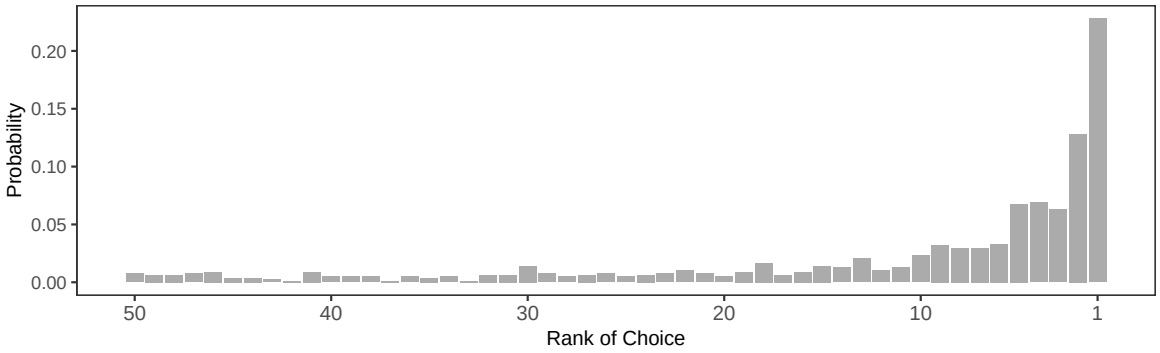
What is clear is that participants were under-sorting, and the cascading down heuristic was the predominant strategy by which they sorted their lists. This systematic under-sorting

has important implications for our evaluation strategy. We cannot simply treat participants' Phase I rankings as ground truth when assessing the quality of LLM-derived preferences. The rankings themselves are incomplete revelations of participants' true preferences, particularly in the middle and lower portions of the lists. This observation motivates our Phase II validation, where we directly elicit participants' preferences between only a few tasks, providing a more reliable benchmark for comparing the LLM-elicited preferences.

### 3.2 Single job application game choices

To complement our understanding of how participants express preferences, we examine their behavior in a more focused elicitation context. In the final step of Phase I, each participant was asked to select a single task in a one-shot single job application, knowing that others might compete for the same task. Figure 5 plots the rank of that chosen task relative to each participant's own final ordering.

**Figure 5:** Rank of participants' single job-game choice compared to their sorted preferences




*Notes: This figure shows how participants' single job-game choice tasks are distributed across their final ordinal rankings. For example, a substantial fraction of participants (23%) selected their absolute favorite (rank 1) as their job-game choice.*

A substantial fraction of participants selected their top-ranked option, with 23% indeed picking their outright favorite, while 70% chose a task within their top ten.

This distribution suggests that participants took the possibility of congestion seriously rather than choosing purely at random or always defaulting to their absolute favorite. More generally this provides evidence that participants were actively completing the elicitation exercises, yet they still failed to fully sort their preference lists in the ranking phase. Fur-

ther evidence of participants' thoughtful engagement can be found in their natural language descriptions of preferences, which we analyze next.

### 3.3 Analysis of natural language descriptions

We now examine the qualitative content of participants' natural language taste descriptions and illustrate what sorts of information they convey. Table 1 presents four examples.

**Table 1:** Four examples of participant's natural language tastes

| | **Likes** | **Dislikes** |
|---|---|---|
| **A** | *"I mostly use my phone for prolific, so preferably tasks that are compatible on mobile. Preferably shorter tasks, under 25 minutes. Multiple choice questions are ideal. Open ended are fine if not too extensive. I don't like tasks that have repetitive questions using different words. But I'm open to mostly any topic of task."* | *"Avoid tasks that require downloading an app, using a camera, or going out of my way to do work on other apps required for the task."* |
| **B** | *"Do not choose any that involve extensive writing or speaking. Also, avoid tasks that involve the use of a microphone or webcam, as I do not have those set up at this time, and i prefer my privacy. I would prefer ones that are quick and pay at least a dollar, but that's more of a preference and not a requirement."* | *"I would avoid ones that take longer than about 20 minutes, despite the incentive looking pleasing, I cannot fully engage in a survey that takes long, or I will lose focus and not finish."* |
| **C** | *"I like to do short preference studies. I prefer to do many shorter ones versus doing longer format ones. I do not like to do a lot of political studies, but do enjoy providing feedback for businesses. I enjoy studies that have lots of different questions, not the same questions formatted differently."* | *"I do not enjoy any studies that take longer than ten minutes. I do not like to do studies that require lots of my personal information."* |
| **D** | *"I enjoy tasks on Prolific that are about 5-10 minutes for at least $.50. I like workplace scenarios, product ratings, political surveys, tasks with other Prolific users. I like most tasks and am not too picky. It is easier for me to use mobile for tasks but I am open to using desktop if a survey seems fun or interesting."* | *"I do not like tasks related to animals and I prefer to avoid them. That is my only preference because I had a task that did not have a warning but had graphic upsetting material of an animal."* |

*Notes: Participant responses to: "Write a paragraph providing guidance to a friend picking Prolific tasks on your behalf for you to do later" and "Is there anything in particular that your friend should avoid?"*

These short paragraphs reveal several types of information that would be difficult to infer from rankings alone. First, participants often specify technical constraints and task formats they prefer or avoid. For example, participant A emphasizes mobile compatibility, while participant B makes clear that microphone or webcam usage is a deal-breaker. Time constraints also feature prominently: some participants are open to longer tasks with higher pay, whereas others cap their session length at ten or twenty minutes, regardless of pay.

Moreover, these statements encode broader categorical preferences. Participant D's aversion to animal-related content stems from a previous distressing experience, and participant C

indicates a strong dislike for political surveys, favoring business-related questionnaires instead. These sorts of "hard constraints" and category-level preferences are not always obvious from an ordinal ranking of 50 tasks—especially if the participant never encounters an animal-related study or only a few political items.

These short paragraphs provide qualitative evidence that even in our relatively limited setting, natural language can convey deeper and more generalizable programmatic information about participants' tastes than a simple ordinal ranking might. In particular, they highlight specific constraints (e.g., microphone requirements) and broader preferences (e.g., aversion to certain topics) that could apply equally to tasks beyond the initial set of 50. Such flexibility is precisely what we seek from free-text taste descriptions: a concise yet adaptable blueprint that captures participants' key requirements and principles, complementing or even superseding their under-sorted rankings in Phase I.

### 3.4 Constructing utilities from natural language tastes

Having illustrated the depth and clear programmatic nature of participants' textual preferences, we now describe the process by which we translate these free-text descriptions into vNM utilities. Our procedure follows the classic approach of eliciting indifference probabilities in a lottery (von Neumann and Morgenstern, 1944). In this procedure, each participant is prompted to consider a hypothetical choice between receiving a particular task $j$ for sure or taking a lottery that yields their most preferred task with probability $p$ and their least preferred task with probability $1 - p$. The value of $p$ that renders the participant indifferent between these two options defines the utility of task $j$ on a $[0, 1]$ scale.

To do this with an LLM, we first identify each participant $i$s favorite task, $r_i^+$, and assign it a utility of 1. Although participants often under-sorted the middle portion of their lists, they consistently placed their genuine top choice at rank 1, making $r_i^+$ a natural anchor for the upper bound of the utility scale. For the lower bound, we prompt an LLM with the participant's free-text description $\mathcal{T}_i$ and ask it to determine which task appears most misaligned with those stated preferences, denoting that worst task by $r_i^-$ and assigning it a

utility of 0.

Once these anchors are in place, we partition the remaining tasks into batches of six. We partition the tasks because the LLM sometimes outputs the incorrect number of utilities when evaluating too many tasks at once. For each task $j$ in a batch, we feed $\mathcal{T}_i$, $r_i^+$, and $r_i^-$ into the LLM and ask it to identify the probability $p_{ij}$ at which the participant would be indifferent between receiving task $j$ for certain and taking the lottery that yields $r_i^+$ with probability $p_{ij}$ and $r_i^-$ with probability $1 - p_{ij}$. The complete prompt is shown in Figure A11 of the Appendix. Repeating this process for all tasks yields the vNM utility function $u_i(j) = p_{ij}$ for each participant $i$. Because every batch is anchored by the same $r_i^+$ and $r_i^-$, we obtain a unified 0-1 utility scale across all tasks.

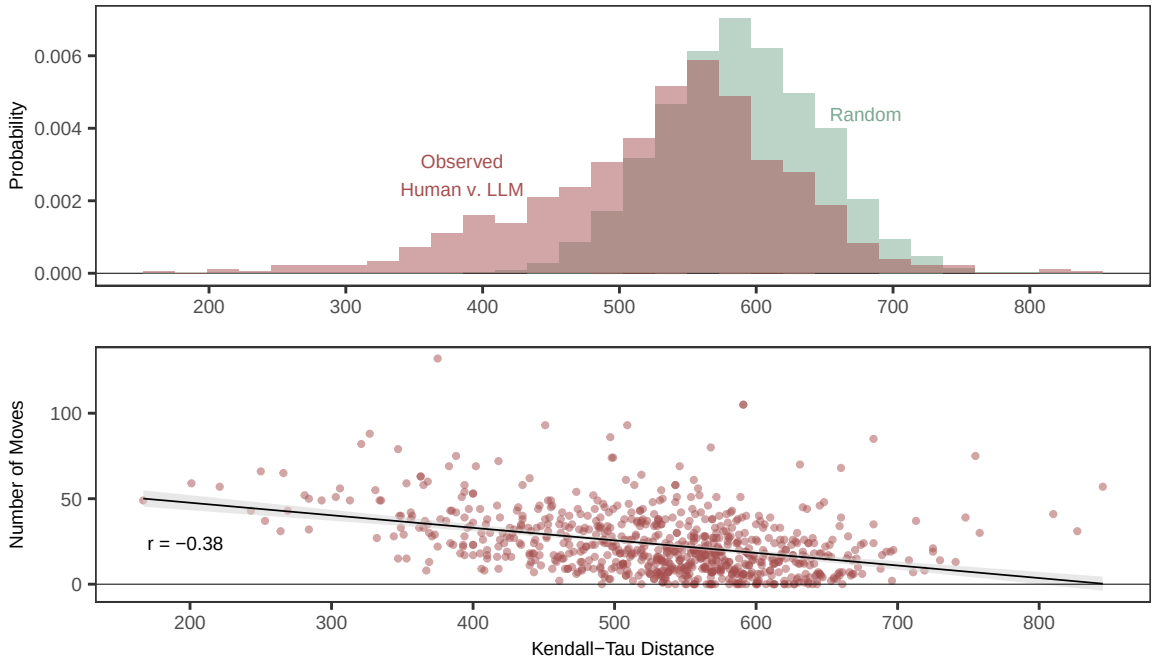
### 3.4.1 Evaluating LLM-elicited utilities

To quantify a preliminary measure of how well our LLM-elicited utilities capture human preferences, we compare two sets of rankings. First, we calculate the Kendall-$\tau$ distance, the number of pairwise disagreements between ordered lists, between each participant's rankings, and the rankings implied by the vNM utilities elicited from the LLM.[5] Each participant's first choice $r_i^+$ is excluded from both rankings as these were identical by design. Next, to establish a baseline for comparison, we take a single list of 49 items and generate 10,000 random permutations, computing the Kendall-$\tau$ distance for each permutation against the original ordering.

The top panel of Figure 6 shows these two distributions—the actual distances between human and LLM rankings (in red) and the distribution of distances from the random permutations (in green). The human-LLM distribution is shifted far closer to 0 relative to the random permutations. This suggests that our utility elicitation method successfully extracts meaningful preference information from users' natural language taste descriptions. Although there is clear room for improvement, such a result is impressive given that we did little prompt engineering and optimization to elicit the LLMs' utilities.

[5]Kendall-$\tau$ distance is distinct from the minimal insertion-sort metric, which captures how many insertion steps are needed to transform one ordering into another.

**Figure 6:** LLM-derived preferences alignment compared to random permutations and correlation with sorting effort




*Notes: In the top panel, the red distribution shows each participant's Kendall-$\tau$ distance between the LLM-based ranking and their final ranking (excluding their top task). The green distribution corresponds to 10,000 random permutations of 49 items against the same reference ordering. The bottom panel shows how this distance relates to the number of re-ranking moves each participant made (y-axis). Notably, the red distribution in the top panel is a histogram of the bottom panel's x-axis measure.*

Still, this comparison likely *underestimates* how well the LLM captures actual preferences, given that many participants under-sorted their lists (hence our additional Phase II experiment). The bottom panel of Figure 6 offers direct evidence: it shows that participants who devote more effort to re-ranking—that is, those providing more reliable preference data—exhibit significantly smaller Kendall-$\tau$ distances. The same pattern arises when using time spent ranking rather than the number of moves (see Figure A9). Thus, if sorting effort is a plausible proxy for accurately reported preferences, then our measures of the alignment between LLM-derived utilities and human preferences likely are, if anything, an conservative.

# 4 Generating matches

Our analysis now shifts to constructing formal matching outcomes from the LLM-derived preference data and the participant rankings from Phase I. We use this data to simulate three distinct matching mechanisms—the Hylland-Zeckhauser pseudo-market (requiring cardinal preferences), random serial dictatorship (requiring ordinal preferences), and a single job application game (requiring a maximizing choice)—under both high- and low-congestion scenarios. Finally, we compare the resulting matches according to both the LLM-implied utilities and participants' own ordinal rankings.

## 4.1 Simulating different levels of congestion

To model different levels of congestion, we begin by estimating how likely participants are to receive each rank in a typical random serial dictatorship outcome based on their submitted lists. Because participants often undersort the lower portions of their lists, we do not want to give every rank position equal weight—doing so could obscure how competition unfolds at the top. Instead, we run 1,000 random serial dictatorship simulations on random 50-person subsets of our participant pool, always using each individual's submitted list for the seeker side of the market, and observe the final match rank for each participant in each simulation. Stacking these outcomes produces a distribution of how frequently each rank (from 1 to 50) is realized. We then add a small uniform offset for smoothing.

The resulting empirical frequency distribution (see Table A2 in the Appendix) shows that more than 90% of matches occur in the top 15 ranks. This weighting ensures that disagreements in the middle or bottom of the ranking—which many participants barely sorted—do not artificially dilute or inflate how similar participants' preferences appear to be.

Given this distribution, we define a position-based weighting function $w(r)$ equal to the probability of being matched at rank $r$. Let $\sigma$ and $\tau$ be two rankings (permutations) of $n$ items. In the ranking $\sigma$, $\sigma(i)$ denotes the item that occupies position $i$. To measure how different $\sigma$ and $\tau$ are, we sum $w(i) + w(j)$ for each pair of positions $(i, j)$ that is inverted. A pair is inverted if $\sigma(i)$ and $\sigma(j)$ appear in one order under $\sigma$ but in the reverse order under

$\tau$; mathematically,

$$\sigma(i) < \sigma(j) \ \wedge \ \tau(i) > \tau(j) \quad \text{or} \quad \sigma(i) > \sigma(j) \ \wedge \ \tau(i) < \tau(j).$$

Thus, the weighted Kendall-$\tau$ distance is

$$d(\sigma, \tau) \ = \ \frac{\sum_{1 \le i < j \le n} \Big[ \big( w(i) + w(j) \big) \ \cdot \ \mathbf{1}\{\sigma(i) < \sigma(j) \ \wedge \ \tau(i) > \tau(j)\} \Big]}{\sum_{1 \le i < j \le n} \big( w(i) + w(j) \big)} \, .$$

Because $w(r)$ is largest for ranks near the top (where participants are more likely to be matched) and decreases monotonically, disagreements among top-ranked items receive more weight than those among lower-ranked items.

Using this distance measure, we created two market scenarios for each participant: (i) a maximally congested market, with the 49 participants having the most similar preferences according to the weighted Kendall-$\tau$ distance, and (ii) a maximally uncongested market, with the 49 participants having the most different preferences according to the distance metric. We then ran three matching mechanisms—Hylland-Zeckhauser, random serial dictatorship, and a conventional job application game—in both markets for each participant, yielding a dataset of six potential matches per participant.

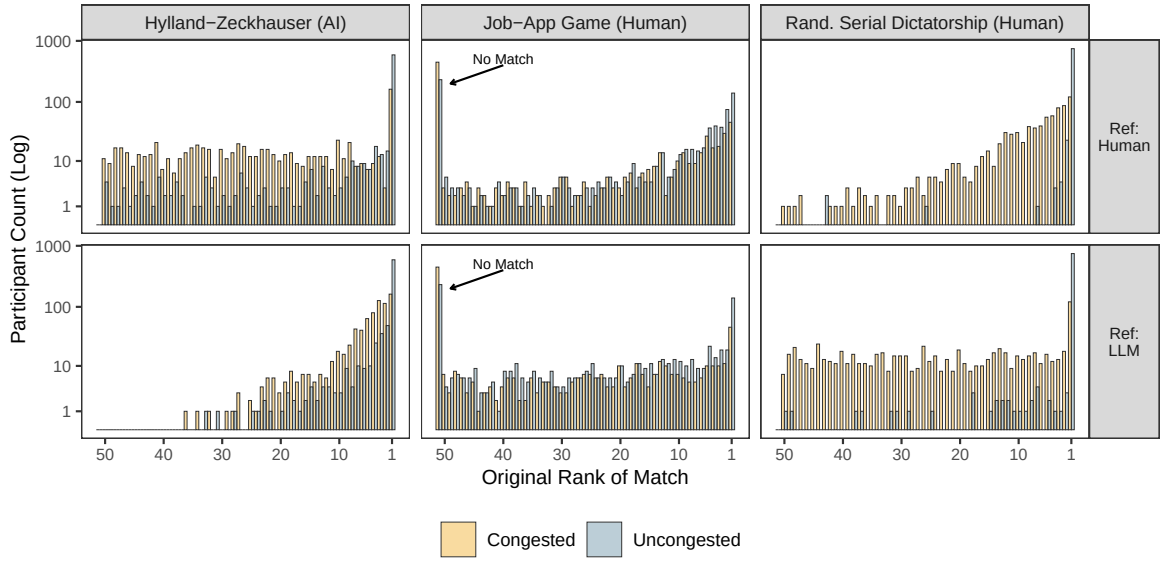
### 4.2 Simulated welfare analyses

Because participants often under-sorted their rankings, we evaluate each mechanism's outcomes from two different perspectives: the participant's own ordinal list and the LLM-derived vNM utilities. Neither perspective constitutes ground truth, but together, they offer a more holistic sense of welfare outcomes.

Figure 7 illustrates the distribution of final matches (for the "focal participant") in each market simulation under random serial dictatorship (random serial dictatorship), Hylland-Zeckhauser, and the single-job-application game. The top row shows matches ranked by participants' own lists (*Ref: Human*), and the bottom row uses LLM-implied utilities (*Ref: LLM*). Yellow bars represent the matches from the maximally congested markets (with the

49 participants with the most similar preferences), while blue bars represent the matches from the maximally uncongested markets (with the 49 participants with the most different preferences).

**Figure 7:** Distribution of matches by original rank and mechanism under varying levels of congestion




*Notes: Histograms show matched task ranks from two perspectives: participant's ordinal list (Ref: Human, top) and LLM-implied vNM utilities (Ref: LLM, bottom). Mechanisms compared: Random Serial Dictatorship (random serial dictatorship), Hylland-Zeckhauser, and single job application game. Yellow bars show congested markets (high preference overlap); blue bars show uncongested markets (low overlap). Rank 1 = top choice; y-axis shows log-scale participant counts.*

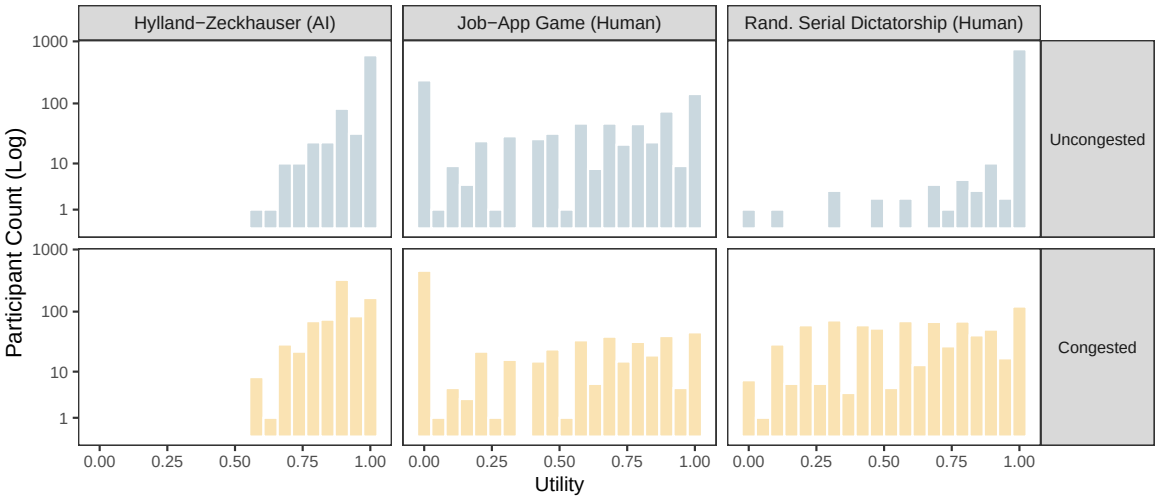
In uncongested markets (blue bars), random serial dictatorship generally places participants in their top choice under either reference, with most participants being matched to rank 1. Hylland-Zeckhauser also tends to place participants well in these uncongested settings, though it performs slightly worse than random serial dictatorship whether viewed from the human or LLM perspective. By contrast, the single-job-application game fares poorly, often leading to unmatched participants or undesirable outcomes

Under congestion (yellow bars), all three mechanisms shift more participants away from their top-ranked tasks in both references. However, Hylland-Zeckhauser appears far less vulnerable to crowding, allocating more participants closer to their top tiers than random serial dictatorship or the single-job-application game. This resilience aligns with the well-

known theoretical result that Pareto-efficient, cardinal mechanisms can help mitigate the externalities of competition for a single popular slot (Hylland and Zeckhauser, 1979).

Turning to cardinal utilities, Figure 8 displays how many participants receive each possible vNM utility (0 to 1), as derived by the LLM, under each mechanism, on a log-scale y-axis. As in Figure 7, the top row corresponds to uncongested markets, and the bottom row to congested markets. In the uncongested scenario, random serial dictatorship and Hylland-Zeckhauser both place most participants at or near utility 1, whereas the single-job-application game yields substantial mass at lower utilities or no match. Under congestion, Hylland-Zeckhauser remains comparatively robust, retaining a significant portion of matches in the higher utility range.

**Figure 8:** Distribution of vNM utilities by mechanism under varying levels of congestion




*Notes: Histograms show the distribution of participants' LLM-derived vNM utilities (0 to 1, log-scale y-axis) under three mechanisms: Random Serial Dictatorship (random serial dictatorship, left), Hylland-Zeckhauser (middle), and single job application game (right). The top row shows uncongested markets (rarely overlapping top choices); bottom row shows congested markets (highly overlapping top choices).*

One potential concern is that our evaluation uses the same LLM-derived utilities that guide the Hylland-Zeckhauser mechanism, creating the possibility of circularity. However, the LLM's strong alignment with participants' top-ranked items suggests these welfare improvements do not simply reflect self-reinforcement. Still, the LLM's preferences are not necessarily perfect, which is why we now turn to a follow-up experiment to provide direct evaluation of the LLM-derived utilities.

# 5 Phase II results

Given the inherent challenges in determining ground truth preferences, both from participants' incomplete sorting and the LLM's derived utilities, we conducted a follow-up validation to provide direct evidence of mechanism performance. In this phase, participants made explicit choices between the matches proposed by different mechanisms and, therefore, different preference elicitation methods. This provides a better ground truth for comparing the performance since participants only need to choose between a few limited options.

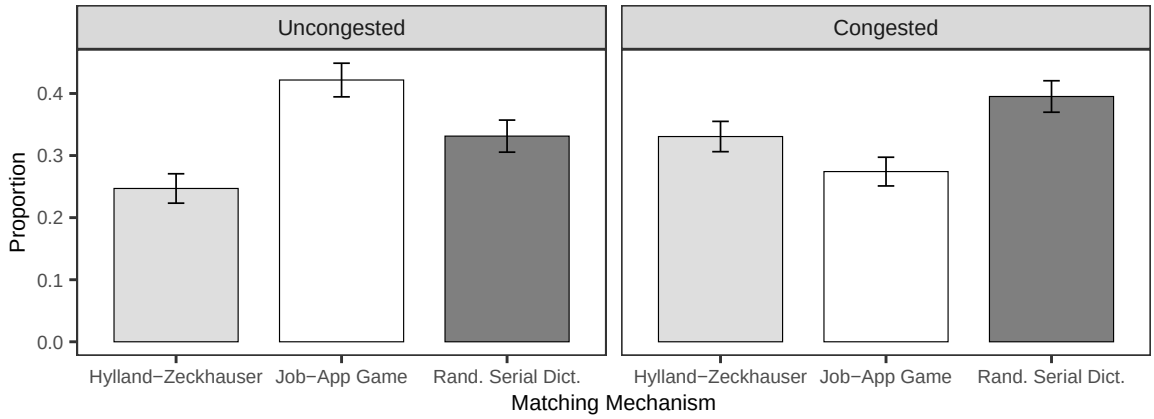
In this follow-up, participants were shown three jobs derived from the simulated matches generated in the previous section based on their submitted lists and the LLM-derived vNM utilities. Participants were randomly assigned to one of two treatments based on the weighted Kendall-$\tau$ distance among their matched peers: a congested group (where preferences overlapped heavily) or an uncongested group (where preferences were relatively dissimilar). From these respective matches, each participant then saw three tasks, one assigned by each mechanism (random serial dictatorship, Hylland-Zeckhauser, and the single job application game), and was asked to pick which they most preferred in that moment.[6]

Figure 9 displays the fraction of participants selecting each mechanism's assigned task under the two conditions. When participants chose matches generated from an uncongested market, they were most likely to pick the job application game-assigned task (white bars), at roughly 45%, followed by random serial dictatorship at about 35%. Fewer participants chose the Hylland-Zeckhauser task in these uncongested settings. By contrast, in the congested condition, random serial dictatorship became the leading choice (dark bars), while the game-assigned option saw its share drop to around 30%. Notably, the share of participants selecting Hylland-Zeckhauser (gray bars) rose substantially under congestion, underscoring that its relative value seems higher precisely when participants' top choices heavily overlap.

Statistically, a within-treatment chi-square goodness-of-fit test (against a uniform distribution of $\frac{1}{3}$ for each mechanism) confirms that the three choices differ significantly in the

[6] In a small fraction of cases, two mechanisms yielded the same assigned task or the game yielded no assignment, which reduced the choice set to two.

**Figure 9:** Proportion choosing each mechanism's match by treatment




*Notes: For each treatment, bars display the fraction choosing the match from random serial dictatorship (dark), Hylland-Zeckhauser (gray), or the single job application game (white). Error bars represent 95% confidence intervals.*

uncongested group ($\chi^2(2) = 15.2, p < 0.001$) and the congested group ($\chi^2(2) = 0.017, p \approx 0.017$). Moreover, a between-treatment chi-square test of independence on the full $2 \times 3$ contingency table yields $\chi^2(2) = 17.28, p < .0001$, indicating that the choice distribution in congested markets is statistically different from that in uncongested markets overall.

Finally, we compare the final chosen matches to the participant's original ranked lists (see Figure A10). In uncongested markets, a large fraction of participants—63%—were matched with their absolute top choice via at least one of the mechanisms. Similarly, 83% of participants in uncongested markets received a top-10 task under some mechanism shown in the final triplet. In congested markets, these shares drop noticeably (to 26% to get one's top choice and 67% in top-10 matches), highlighting how stiff competition can push people away from their most preferred outcomes. Still, many participants in the congested condition ended up with tasks they had originally placed fairly high in their own rankings, particularly under random serial dictatorship and Hylland-Zeckhauser. Thus, for most participants, the final choice among the three mechanism-proposed jobs was not a low-ranked or otherwise obviously undesirable option.

# 6 Discussion

Our findings show that even brief, free-text descriptions of tastes can substantially reduce the cognitive burdens inherent in markets requiring detailed preference data. By translating these descriptions into vNM utilities, LLMs ease the classic accuracy-burden trade-off: participants need not fully sort large sets of items, yet still achieve better-than-random alignment. Furthermore, incorporating vNM utilities proves particularly beneficial in congested settings, where many participants chase overlapping top choices.

A key challenge throughout our study is the elusive notion of ground truth. Participants' self-reported rankings may be incomplete beyond their top few choices, and LLMs—reliant on textual prompts—can fail in unforeseen ways. Our follow-up experiment helps disentangle these issues by providing an unbiased causal comparison of outcomes, but it still offers only a partial view of participants' full preference maps. More extensive validation approaches, such as multi-stage elicitation or repeated refinements in larger real-world markets, could better illuminate how accurately LLM-derived utilities capture complex trade-offs that participants themselves struggle to articulate.

One limitation of our Phase II experiment is that participants often encountered at least one or two tasks they had ranked highly. This limited how strongly they had to weigh choices among less-preferred tasks. Yet in many real-world settings—where mismatches and competition are common—participants end up choosing among second- or even third-tier options. Since our results show that the lower portion of participants' lists is the least sorted, it is exactly in these "middle-ground" trade-offs that we might expect LLM-driven elicitation to provide the largest marginal improvement. Future work could thus explore experimental designs or natural market conditions that push participants deeper into their lists, enabling a clearer test of whether LLM utilities outperform hastily formed human ranks for items that are "acceptable but not ideal."

Indeed, recent work suggests that as language models improve, they will become increasingly intuitive for participants to use (Jahani et al., 2024). We expect humans will adapt in tandem, refining prompt-based descriptions to specify their preferences more com-

prehensively and reducing reliance on cumbersome checklists or partial rankings. Although challenges regarding reliability, interpretability, and potential strategic manipulation remain, prompt refinements, user-interface design, and improved alignment techniques offer plausible paths to greater robustness.

The direct applications of this method could be transformative for complex, real-world matching markets. Consider the example of traveling nurses, each choosing among thousands of short-term hospital contracts nationwide. This market is so complex that nurses often use intermediary agencies rather than a single, large-scale matching system. However, one could imagine a scenario in which a nurse briefly describes his tastes: *"I prefer larger urban hospitals that serve under-resourced patients, am willing to work night shifts if the pay compensates, and need housing stipends that cover most living expenses."* An LLM might then use repeated lotteries to calibrate how intensely he prefers, say, Hospital B (a large public hospital in a major city) over Hospital A (a smaller private facility in a rural area). Accumulating such cardinal utilities across dozens or hundreds of openings, the system could produce a precise ranking that respects the user's trade-offs. If hospitals similarly elicited their own preferences over nurses, a matching mechanism could then provide allocations that work better for both parties. These kinds of large and dynamic settings—including dating platforms (Hitsch et al., 2010), job marketplaces, and beyond—stand to benefit greatly from offloading preference discovery to LLMs, as human participants regularly face cognitive overload.

# 7 Conclusion

This paper demonstrates that large language models can substantially reduce the cognitive burden of reporting detailed preferences in matching environments. By allowing participants to specify a concise paragraph of textual tastes, our approach generates preferences that can be used to implement and evaluate superior market designs at scale. In particular, our results show that when participants face limited overlap in their top choices, simple ordinal mechanisms already work well; however, as overlap intensifies—and thus congestion rises—LLM-derived utilities enable mechanisms like Hylland-Zeckhauser to achieve better

allocations and higher overall welfare.

We find that participants' preferences correlate strongly with LLM-implied rankings, and this alignment improves further as participants spend more effort refining their own preference lists. At the same time, establishing a perfect ground truth is challenging. To address this, we conducted a follow-up experiment in which participants directly chose among tasks generated by different elicitation methods. Their choices revealed that the LLM-based matches were notably favored in congested scenarios, providing direct causal evidence that participants prefer these assignments over simpler baselines.

Overall, our findings suggest that LLM-based preference elicitation can help bridge the gap between elegant economic theory and practical market design. Allowing participants to describe their tastes in a concise, program-like format spares them from manually ranking large sets of items—a process our experiment shows to be both time-consuming and error-prone. By harnessing these artificially derived preferences, mechanism designers can more effectively manage congestion and offer participants matches that better align with their actual interests. In this sense, our work is an initial step toward realizing the long-recognized promise of market mechanisms that rely on detailed preference data at scale.

## References

**Akerlof, George A. and Janet L. Yellen**, "Can Small Deviations from Rationality Make Significant Differences to Economic Equilibria?," *The American Economic Review*, 1985, *75* (4), 708–720.

**Aziz, Haris, Felix Brandt, and Markus Brill**, "The computational complexity of random serial dictatorship," *Economics Letters*, 2013, *121* (3), 341–345.

**Brero, Gianluca, Benjamin Lubin, and Sven Seuken**, "Machine Learning-powered Iterative Combinatorial Auctions," 2021.

**Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg,**

**Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang**, "Sparks of Artificial General Intelligence: Early experiments with GPT-4," 2023.

**Budish, Eric**, "The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes," *Journal of Political Economy*, 2011, *119* (6), 1061–1103.

**Chapman, Jonathan, Pietro Ortoleva, Erik Snowberg, Leeat Yariv, and Colin Camerer**, "Reassessing Qualitative Self-Assessments and Experimental Validation," Working Paper 33520, National Bureau of Economic Research February 2025.

**Cowen, Tyler**, "Are all tastes constant and identical?: A critique of Stigler and Becker," *Journal of Economic Behavior & Organization*, 1989, *11* (1), 127–135.

**Enke, Benjamin**, "The Cognitive Turn in Behavioral Economics," July 2024. Working paper.

**Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde**, "Global Evidence on Economic Preferences," *The Quarterly Journal of Economics*, 05 2018, *133* (4), 1645–1692.

**_ , _ , _ , David Huffman, and Uwe Sunde**, "The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences," *Management Science*, 2022, *69* (4), 1935–1950.

**Fradkin, Andrey, Monica Bhole, and John J. Horton**, "Competition Avoidance vs Herding in Job Search: Evidence from Large-scale Field Experiments on an Online Job Board," *Management Science*, 2024.

**Hitsch, Gunter J., Ali Hortaçsu, and Dan Ariely**, "Matching and Sorting in Online Dating," *American Economic Review*, March 2010, *100* (1), 130–63.

**Horton, John J**, "Large language models as simulated economic agents: What can we learn from homo silicus?," Technical Report, National Bureau of Economic Research 2023.

**Horton, John J., David G. Rand, and Richard J. Zeckhauser**, "The online laboratory: conducting experiments in a real labor market," *Experimental Economics*, 2011, *14*, 399–425.

**Horton, John J, Shoshana Vasserman, and Mitchell Watt**, "Reducing Congestion in Labor Markets: A Case Study in Simple Market Design," Technical Report, National Bureau of Economic Research 2024.

**Huang, David, Francisco Marmolejo-Cossío, Edwin Lock, and David Parkes**, "Accelerated Preference Elicitation with LLM-Based Proxies," 2025.

**Hylland, Aanund and Richard Zeckhauser**, "The efficient allocation of individuals to positions," *Journal of Political economy*, 1979, *87* (2), 293–314.

**Jahani, Eaman, Benjamin S. Manning, Joe Zhang, Hong-Yi TuYe, Mohammed Alsobay, Christos Nicolaides, Siddharth Suri, and David Holtz**, "As Generative Models Improve, We Must Adapt Our Prompts," 2024.

**Kim, Jeongbin, Matthew Kovach, Kyu-Min Lee, Euncheol Shin, and Hector Tzavellas**, "Learning to be Homo Economicus: Can an LLM Learn Preferences from Choice," 2024.

**Li, Belinda Z., Alex Tamkin, Noah Goodman, and Jacob Andreas**, "Eliciting Human Preferences with Language Models," 2023.

**Lira, Benjamin, Margo Gardner, Abigail Quirk, Cathlyn Stone, Arjun Rao, Lyle Ungar, Stephen Hutt, Louis Hickman, Sidney K. D'Mello, and Angela L. Duckworth**, "Using artificial intelligence to assess personal qualities in college admissions," *Science Advances*, 2023, *9* (41), eadg9405.

**Manning, Benjamin S., Kehang Zhu, and John J. Horton**, "Automated Social Science: Language Models as Scientist and Subjects," Technical Report, NBER 3 2024. Accessed: 2024-03-12.

**Oprea, Ryan**, "Decisions under Risk Are Decisions under Complexity," *American Economic Review*, 2024. Available online at `https://www.aeaweb.org/journals/aer/forthcoming`.

**Schensted, C.**, "Longest Increasing and Decreasing Subsequences," *Canadian Journal of Mathematics*, 1961, *13*, 179–191.

**Soumalias, Ermis, Yanchen Jiang, Kehang Zhu, Michael Curry, Sven Seuken, and David C. Parkes**, "LLM-Powered Preference Elicitation in Combinatorial Assignment," 2025.

**Stigler, George J. and Gary S. Becker**, "De Gustibus Non Est Disputandum," *The American Economic Review*, 1977, *67* (2), 76–90. Accessed 18 Feb. 2025.

**Tessler, Michael Henry, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes et al.**, "AI can help humans find common ground in democratic deliberation," *Science*, 2024, *386* (6719), eadq2852.

**Vafa, Keyon, Ashesh Rambachan, and Sendhil Mullainathan**, "Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function," 2024.

**Vazirani, Vijay V and Mihalis Yannakakis**, "Computational complexity of the Hylland-Zeckhauser scheme for one-sided matching markets," *arXiv preprint arXiv:2004.01348*, 2020.

**von Neumann, John and Oskar Morgenstern**, *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press, 1944.

# A Student-project matching

One of the authors taught an MBA course in which teams of students collaborated on real generative AI projects for technology companies. Each team was required to rank 31 possible company sponsors using the exact same interface as the one used in the Prolific study. Using the company and student profiles, we asked an LLM to utilities on behalf of the companies over the students. We then used a student-proposing deferred acceptance to match each team to exactly one company project for the duration of the semester. The stakes were high: students knew their semester-long project would be determined by this ranking and they would have to spend months working on deliverables specified by their matched company, ultimately presenting final outputs that integrated current generative AI methods.

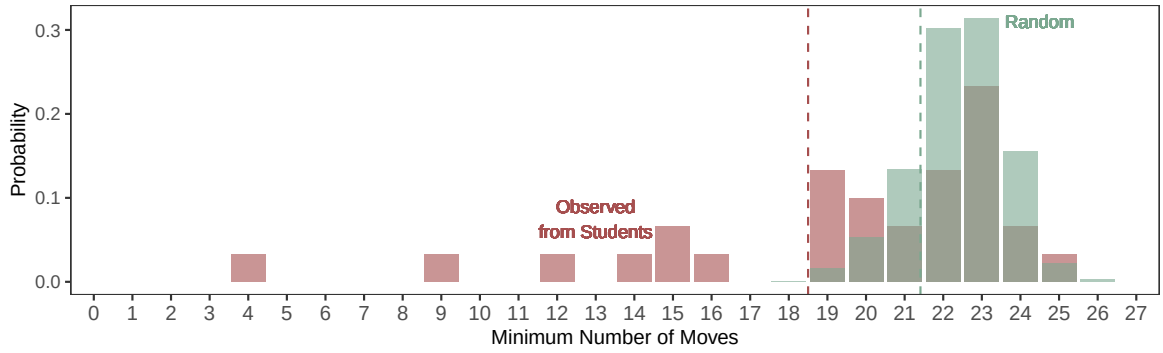
Figure A1 examines the degree to which students sorted their lists beyond their top few choices. This figure is identical to Figure 4 in the main text, except that it uses the student-project matching data instead of the Prolific participant data. We compare the minimal number of insertion moves needed to transform the random initial ordering of projects into each team's final, submitted ranking (in red) against what we would observe for random permutations of 31 items (in green). Despite the tangible academic stakes—students knew their semester-long project would be determined by this ranking—the distribution of minimal moves is well to the left of the random benchmark, indicating that many students under-sorted their lists.

While less severe, these results mirror the main paper's findings with Prolific participants. Even when participants face nontrivial stakes and have time to deliberate, their final reported rankings often remain only partially sorted. This observation further underscores the difficulties of generating fully accurate, fine-grained ordinal preference lists.

# B Additional figures and tables

## B.1 Instructions and survey interface

**Figure A1:** Distribution of minimal insertion moves to achieve students' final rankings vs. random permutations




*Notes: The red distribution shows the actual minimal insertion moves needed to turn each student's randomly ordered projects into their final ranking, while the green distribution shows the same statistic for purely random permutations of 31 projects. The dashed vertical lines represent the average number of moves in each distribution.*

**Figure A2:** GPT-4o's response to the example of a participant's fruit preferences

You are a person who loves crisp textures and tart flavors but avoids fruits with pits. Rank the following 4 fruits in order of your favorite to least favorite: Banana, Cherry, Apple, and Date.

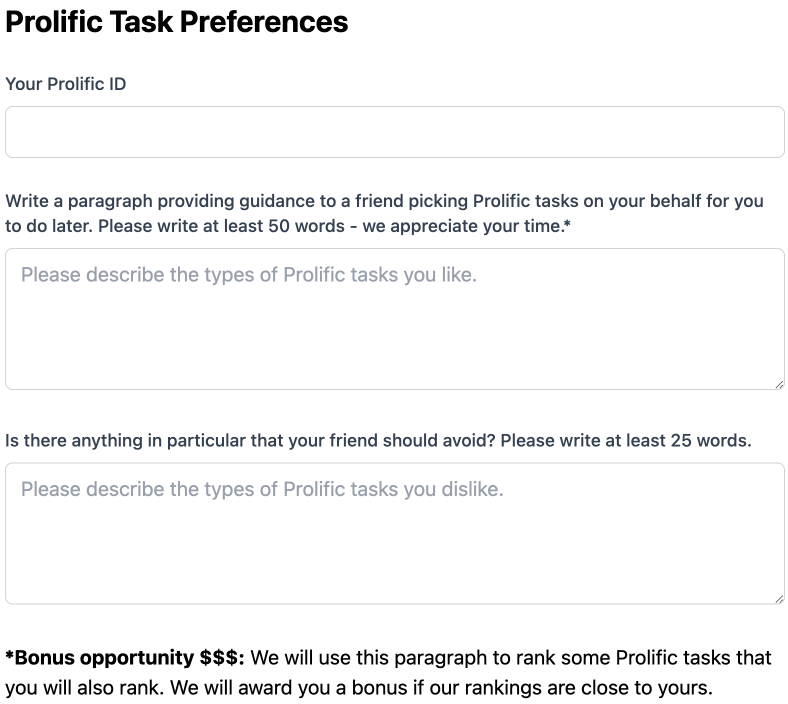
Given that I love **crisp textures** and **tart flavors** but **avoid fruits with pits**, I would rank the fruits as follows:

1. **Apple** – Crisp texture and can be tart (e.g., Granny Smith). No pits, just a core. Perfect!
2. **Banana** – Not crisp, but also not a pitted fruit. Mildly sweet, so not my ideal flavor, but acceptable.
3. **Cherry** – Tartness is great, but it has a pit, which I avoid.
4. **Date** – Soft texture and extremely sweet, plus it has a pit. Least appealing.

So, my ranking from favorite to least favorite is: **Apple > Banana > Cherry > Date.**

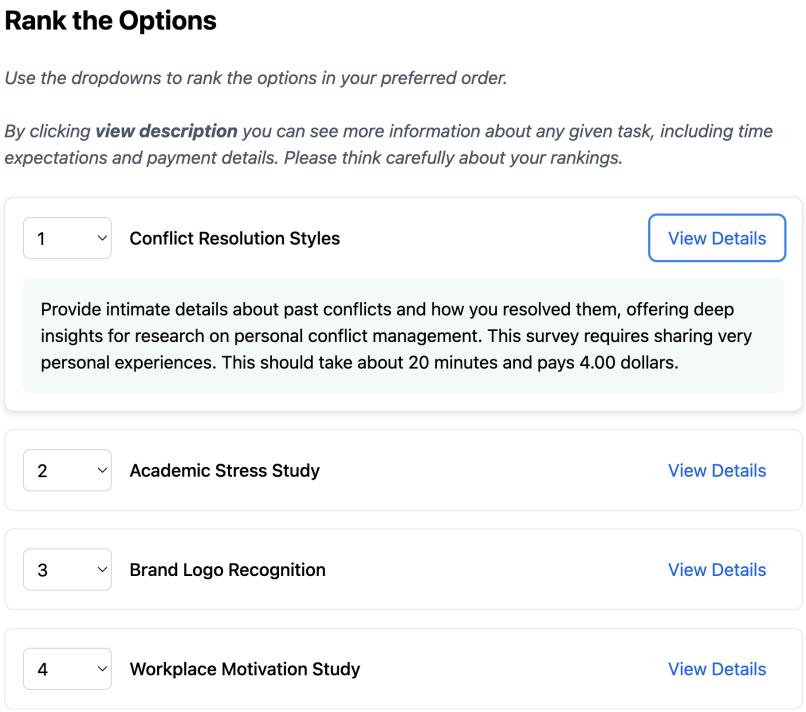
*Notes: The order of fruits is what one would plausibly expect from a participant who likes crisp textures and tart flavors but avoids fruits with pits.*

**Figure A3:** Instructions for natural language preference elicitation



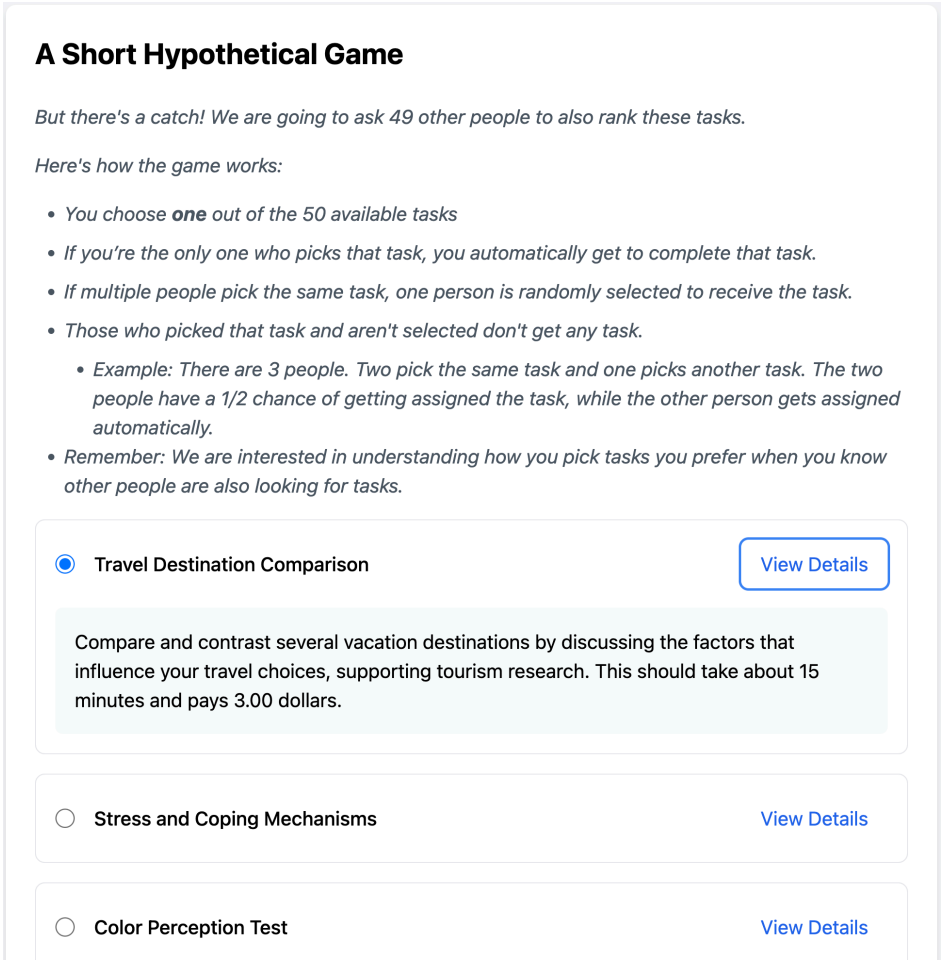

*Notes: Screenshot showing how participants were asked to describe their task preferences in natural language format.*

**Figure A4:** Page to rank tasks



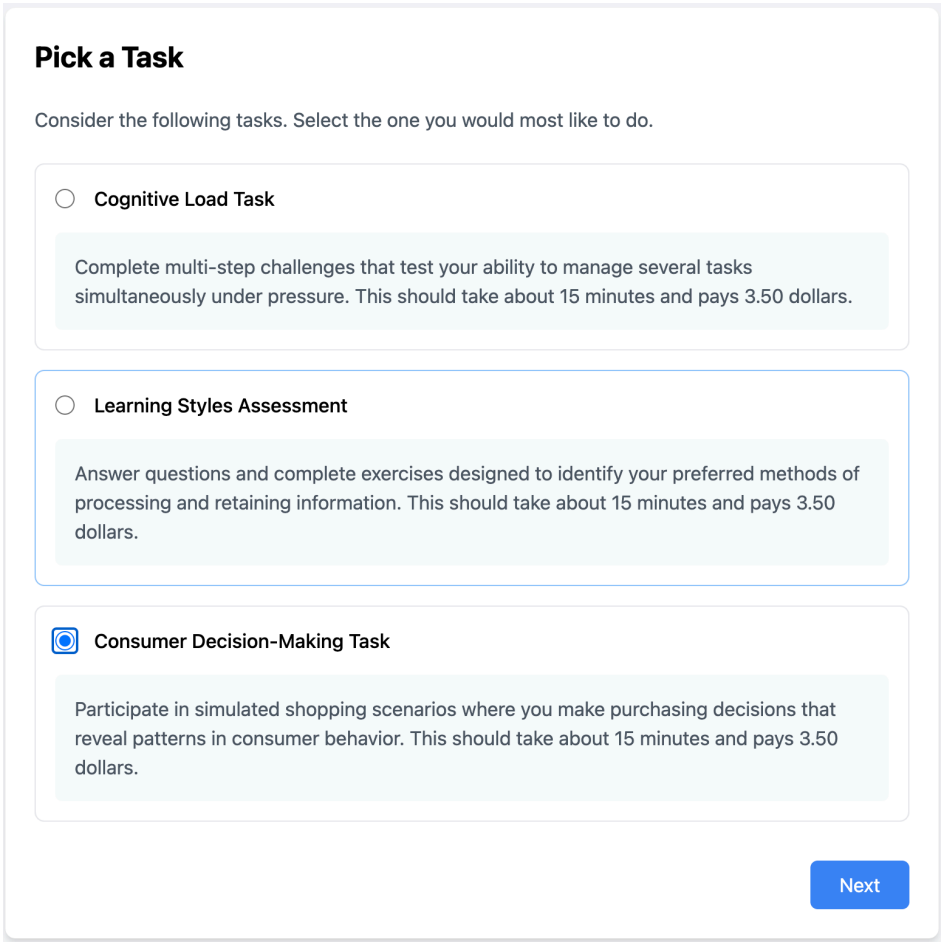

*Notes: Screenshot showing the page where participants ranked tasks.*

**Figure A5:** Instructions for job application game choice




*Notes: Screenshot showing how participants were asked to select a single task for the job application game portion of the experiment.*

**Figure A6:** Follow-up survey interface




*Notes: Screenshot of Phase II follow-up survey where participants evaluated matches generated by different mechanisms.*

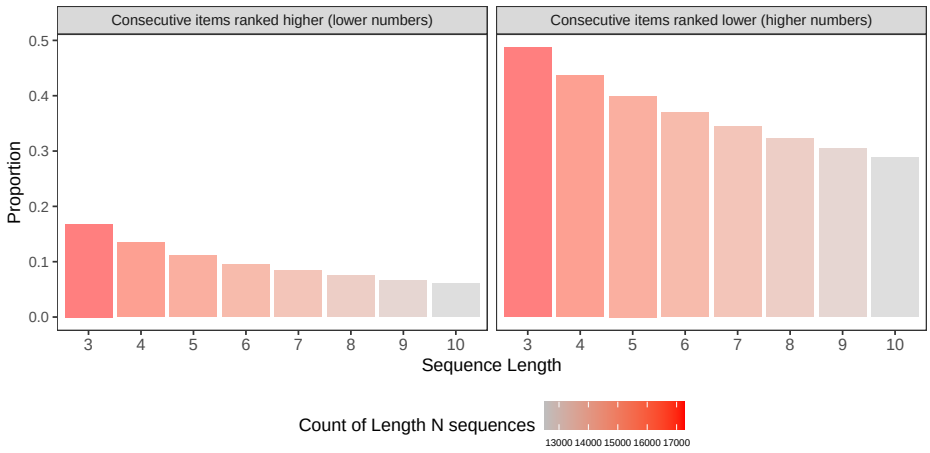
## B.2 Additional analysis

**Table A1:** Regression of the destination of the rank move on the move number

| | *Rank Moved To* | |
|---|---|---|
| | (1) | (2) |
| *Move Number* | 0.230* | 0.264* |
| | (0.007) | (0.026) |
| *Constant* | 15.742* | |
| | (0.173) | |
| Participant FE | No | Yes |
| Clustered SE | No | Yes |
| Observations | 18,756 | 18,756 |
| $R^2$ | 0.055 | 0.243 |

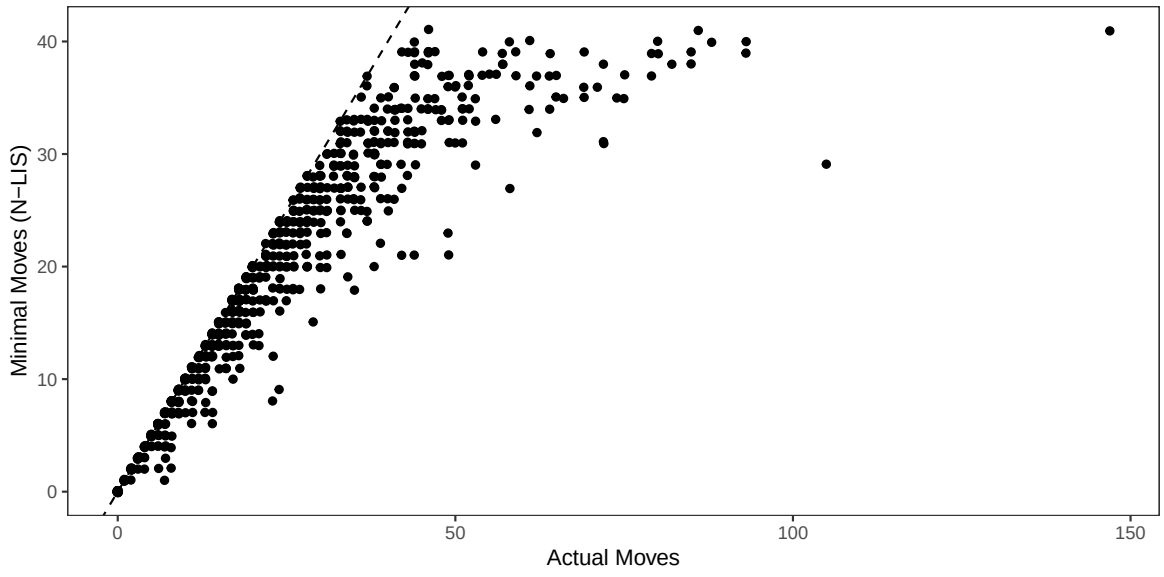
*Notes: This table shows the regression of the destination of the rank move on the move number. For example, the coefficient on move number can be interpreted as the following: on average, participants rank choices 0.230 ranks lower (higher numeric ranking) than the destination rank of the immediately previous task moved. Significance Indicator: *p<0.001.*

**Figure A7:** Proportion of moves of length N where participants monotonically ranked the next item lower than the previous item.



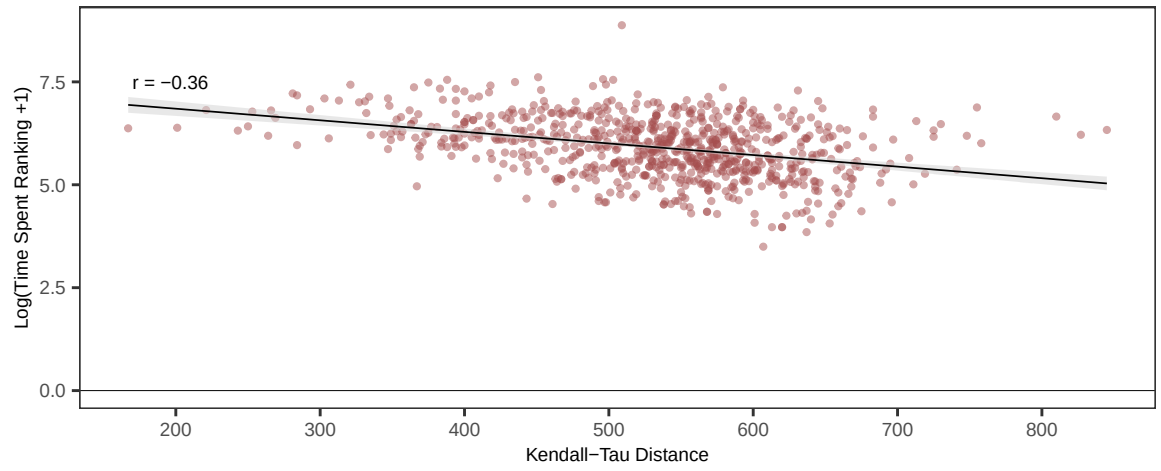

*Notes:*

**Figure A8:** Comparison of the minimum number of reranking steps participants could have taken to move from the initial ranking to the final ranking versus the number of steps they took.



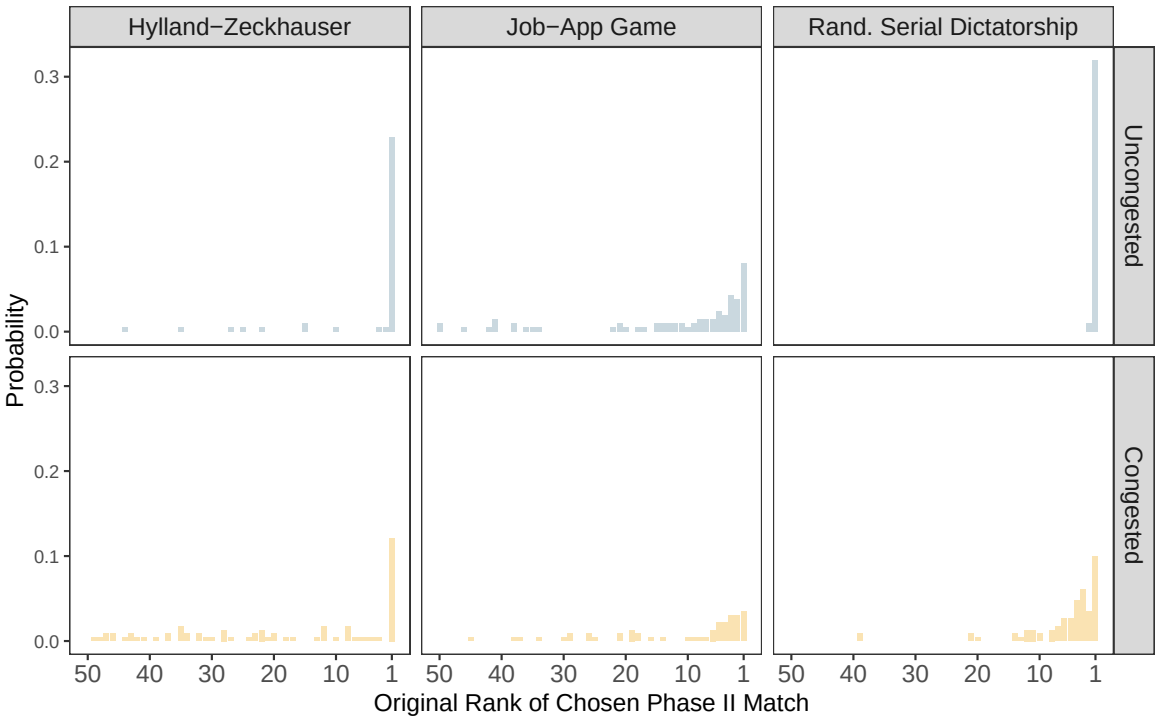

*Notes:*

**Figure A9:** Distribution of Kendall-$\tau$ distances between human and LLM rankings by the log time spent reranking.




*Notes:*

**Figure A10:** Density of participants' original rank of their chosen match by treatment




*Notes:*

**Table A2:** Rank Probabilities

| Rank | Weight | Rank | Weight | Rank | Weight |
|---|---|---|---|---|---|
| 1 | 0.1691 | 18 | 0.0059 | 35 | 0.0019 |
| 2 | 0.1437 | 19 | 0.0051 | 36 | 0.0019 |
| 3 | 0.1183 | 20 | 0.0051 | 37 | 0.0019 |
| 4 | 0.0952 | 21 | 0.0044 | 38 | 0.0019 |
| 5 | 0.0772 | 22 | 0.0035 | 39 | 0.0019 |
| 6 | 0.0612 | 23 | 0.0036 | 40 | 0.0019 |
| 7 | 0.0530 | 24 | 0.0029 | 41 | 0.0019 |
| 8 | 0.0416 | 25 | 0.0028 | 42 | 0.0019 |
| 9 | 0.0329 | 26 | 0.0026 | 43 | 0.0019 |
| 10 | 0.0283 | 27 | 0.0025 | 44 | 0.0018 |
| 11 | 0.0221 | 28 | 0.0021 | 45 | 0.0018 |
| 12 | 0.0187 | 29 | 0.0023 | 46 | 0.0018 |
| 13 | 0.0154 | 30 | 0.0022 | 47 | 0.0018 |
| 14 | 0.0129 | 31 | 0.0021 | 48 | 0.0018 |
| 15 | 0.0106 | 32 | 0.0021 | 49 | 0.0018 |
| 16 | 0.0091 | 33 | 0.0020 | 50 | 0.0018 |
| 17 | 0.0078 | 34 | 0.0019 | | |

*Notes: Probabilities derived from 1,000 Gale-Shapley simulations on random 50-person subsets, using participants' submitted preference lists. Values represent the frequency distribution of match ranks, with a small uniform offset added for smoothing.*

## B.3 Prompts and task descriptions

**Figure A11:** Prompt to elicit utilities

{{persona}} You have the following preferences: {{preferences}} You have already decided that the following task (with its description) is the one you most prefer to do: {{first_choice_job_dict}} You have already decided that the following task is the one you least prefer to do: {{last_choice_job_dict}} Your task is to rate the following tasks on a scale from 0 to 1. Here are the tasks with their descriptions: {{job_list_1_dict}} For each task in the list, assign a probability p between 0 and 1 that represents this scenario: Imagine you could either: A) Do this task for sure, or B) Take a chance where you get your most preferred task {{first_choice_job}} with probability p, or your least preferred task {{last_choice_job}} with probability (1-p) Choose the probability p so that options A and B feel equally appealing to you. Note that for some of the tasks, you may view them very similar (or very different). Based on your preferences, you should use these probabilities to reflect the strength of your ordering. Maybe all of the tasks are quite close to p=1, or maybe they are all quite close to p=0. Or maybe they are evenly distribution, or maybe they are clumped together. Say for example, that you like 3 tasks all equally and close to as much as the favorite, you might rank these 0.9, .9, .9. But if you like 3 tasks very differently (that aren't the first and last), you might rank them 0.5, .9, .10. Or maybe you like them fairly equally, but do have slight preferences, so you might rank them 0.83, .81, .78. Your response should simple be a list of the probabilities (that you decide) corresponding to the order of your rankings. These should include the first and last probabilities (1 and 0) and the middle 8 probabilies for the other tasks. For example, if the list is [{{first_choice_job}}, {{last_choice_job}}, task_1, task_2, task_3, task_4, task_5, task_6, task_7, task_8] your response should be: [1, 0, p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8] since the most preferred task has probability 1, the least preferred has probability 0 (which are given to you), and you have to generate the remaining probabilities yourself. Do not include the task names or any other text in your response. Also make sure you have 10 probabilities in your response - one for each of the 8 tasks plus the first and last.List the probabilities in must align with the order of the tasks: {{job_list_1_to_rank}}, but their relative order should be based on your preferences.

*Notes:*

The following is a list of the tasks that were used in the study:

1. Cognitive Memory Study—In this task, you'll complete a series of memory exercises that test your ability to recall words and images. This should take about 15 minutes and pays 3.50 dollars.
2. Reaction Time Experiment—Participants will respond to rapidly changing visual cues to help researchers analyze reaction speeds under time pressure. This should take about 10 minutes and pays 2.50 dollars.
3. Consumer Behavior Survey—Share insights into your shopping habits and product preferences through a mix of multiple-choice and open-ended questions. This should take about 20 minutes and pays 4.00 dollars.

4. Visual Perception Task—Examine a series of images to spot subtle differences, helping to measure the accuracy of your visual processing skills. This should take about 15 minutes and pays 3.50 dollars.
5. Social Decision-Making Experiment—Engage in interactive scenarios where you make choices that influence outcomes in simulated social situations. This should take about 20 minutes and pays 4.00 dollars.
6. Online Interview Study—Participate in a structured online interview where you share your thoughts on current social trends and personal experiences. This should take about 30 minutes and pays 6.00 dollars.
7. Language Comprehension Task—Read brief passages and answer related questions to assess your understanding and interpretation of written content. This should take about 15 minutes and pays 3.50 dollars.
8. Risk Assessment Study—Make decisions in simulated risky scenarios to help researchers evaluate how people manage uncertainty. This should take about 10 minutes and pays 2.50 dollars.
9. Emotional Regulation Survey—Answer questions about your strategies for managing emotions in everyday situations, contributing to mental health research. This should take about 20 minutes and pays 4.00 dollars.
10. Moral Dilemma Task—Work through challenging ethical scenarios that require balancing competing values and principles. This should take about 15 minutes and pays 3.50 dollars.
11. Attention Span Experiment—Complete a series of tasks designed to measure how well you maintain focus during brief periods of activity. This should take about 10 minutes and pays 2.50 dollars.
12. Implicit Bias Survey—Participate in exercises that reveal subconscious biases through reaction-time tasks and reflective questions. This should take about 20 minutes and pays 4.00 dollars.
13. Learning and Memory Task—Engage with interactive challenges that test how quickly and accurately you can absorb and recall new information. This should take about 15 minutes and pays 3.50 dollars.
14. Problem-Solving Survey—Solve puzzles and answer questions that highlight your analytical and critical thinking skills. This should take about 20 minutes and pays 4.00 dollars.
15. Social Media Behavior Survey—Share your online habits and social media engagement to help researchers understand digital communication patterns. This should take about 20 minutes and pays 4.00 dollars.
16. Face Recognition Task—View and later identify a series of faces to assess your memory and attention to facial details. This should take about 10 minutes and pays 2.50 dollars.
17. Trust and Cooperation Experiment—Participate in interactive games that simulate trust-based decisions and measure cooperative behavior in groups. This should take about 15 minutes and pays 3.50 dollars.
18. Stress Assessment Survey—Reflect on your daily stressors and coping mechanisms through targeted questions aimed at mental health research. This should take about 20 minutes and pays 4.00 dollars.
19. Cognitive Load Task—Complete multi-step challenges that test your ability to manage several tasks simultaneously under pressure. This should take about 15 minutes and pays 3.50 dollars.
20. Environmental Attitudes Survey—Share your opinions on environmental issues and sustainability practices to support ecological research efforts. This should take about 20 minutes and pays 4.00 dollars.
21. Time Perception Task—Estimate time intervals in experiments that explore how accurately you perceive the passage of time. This should take about 10 minutes and pays 2.50 dollars.

22. Humor Appreciation Study—Rate and comment on various humorous scenarios and jokes to help researchers understand what factors contribute to a sense of humor. This should take about 15 minutes and pays 3.50 dollars.

23. Political Attitudes Survey—Express your views on a range of political topics and current events, providing valuable data on public sentiment. This should take about 20 minutes and pays 4.00 dollars.

24. Learning Styles Assessment—Answer questions and complete exercises designed to identify your preferred methods of processing and retaining information. This should take about 15 minutes and pays 3.50 dollars.

25. Attentional Blink Task—Engage in rapid visual tasks that test the brief period during which your brain is momentarily unable to register new stimuli. This should take about 10 minutes and pays 2.50 dollars.

26. Empathy and Compassion Survey—Reflect on your personal experiences and emotional responses to gauge your levels of empathy and compassion. This should take about 20 minutes and pays 4.00 dollars.

27. Online Community Behavior Study—Discuss your participation in online forums and social networks to help researchers understand digital community dynamics. This should take about 15 minutes and pays 3.50 dollars.

28. Creative Writing Task—Compose a short narrative based on provided prompts that encourage creative expression and storytelling. This should take about 45 minutes and pays 9.00 dollars.

29. Digital Game Study—Play a brief, specially designed game that challenges your decision-making and strategic thinking skills. This should take about 5 minutes and pays 1.00 dollars.

30. User Interface Feedback—Test a new app interface by performing guided tasks and providing feedback on its usability and design. This should take about 20 minutes and pays 4.00 dollars.

31. Product Testing Survey—Interact with a new product online and complete a survey detailing your user experience and overall impressions. This should take about 15 minutes and pays 3.50 dollars.

32. Website Usability Study—Navigate a prototype website and share your insights on its design, functionality, and ease of use. This should take about 20 minutes and pays 4.00 dollars.

33. Work-Life Balance Survey—Reflect on your daily routines and strategies for managing work and personal life, contributing to lifestyle research. This should take about 20 minutes and pays 4.00 dollars.

34. Job Satisfaction Survey—Discuss your experiences and satisfaction in your current or past employment, offering valuable insights into workplace dynamics. This should take about 20 minutes and pays 4.00 dollars.

35. Cultural Identity Task—Explore aspects of your cultural background through creative prompts and reflective questions that uncover personal identity. This should take about 15 minutes and pays 3.50 dollars.

36. Personality Inventory—Answer a comprehensive set of questions that map out your personality traits and behavioral tendencies. This should take about 20 minutes and pays 4.00 dollars.

37. Ethical Dilemma Task—Work through thought-provoking scenarios that challenge your moral reasoning and ethical decision-making processes. This should take about 15 minutes and pays 3.50 dollars.

38. Economic Decision-Making Study—Participate in simulated market scenarios where you assess risks and make choices under conditions of uncertainty. This should take about 15 minutes and pays 3.50 dollars.

39. Mindfulness Meditation Survey—Follow a guided meditation session and then complete a survey about your experience with relaxation and focus. This should take about 20 minutes and pays 4.00 dollars.

40. Virtual Reality Experience Feedback—Experience a short virtual reality simulation using your own equipment and provide detailed feedback on the immersive experience. This should take about 40 minutes and pays 9.50 dollars.

41. Memory Recall Task—Engage in exercises that test your ability to remember sequences of numbers and words, contributing to cognitive research. This should take about 15 minutes and pays 3.50 dollars.

42. Dream Analysis Survey—Reflect on your recent dreams and complete a survey that explores recurring themes and emotional tones. This should take about 20 minutes and pays 4.00 dollars.

43. Navigation Task—Plan routes and make directional decisions in a map-based exercise designed to assess your spatial awareness. This should take about 35 minutes and pays 7.50 dollars.

44. Creative Problem-Solving Task—Tackle a series of puzzles and hypothetical challenges that require innovative solutions and creative thinking strategies. This should take about 15 minutes and pays 3.50 dollars.

45. Digital Literacy Survey—Share your experiences with various digital tools and platforms through a survey that examines your overall comfort with technology. This should take about 20 minutes and pays 4.00 dollars.

46. Online Learning Feedback Study—Provide feedback on digital course materials and instructional methods to help improve online learning experiences. This should take about 15 minutes and pays 3.50 dollars.

47. Technology Usage Survey—Detail your daily interactions with smartphones, computers, and other devices, offering insights into modern tech habits. This should take about 20 minutes and pays 4.00 dollars.

48. Consumer Decision-Making Task—Participate in simulated shopping scenarios where you make purchasing decisions that reveal patterns in consumer behavior. This should take about 15 minutes and pays 3.50 dollars.

49. Novelty Seeking Survey—Answer questions about your willingness to try new experiences and explore unfamiliar situations, contributing to research on personality traits. This should take about 20 minutes and pays 4.00 dollars.

50. Health Behavior Survey—Share details about your daily health routines, dietary habits, and exercise practices to support public health research. This should take about 20 minutes and pays 4.00 dollars.