

Human Bottlenecks in AI-Augmented Choice: Evidence from Portfolio Choice *

Geoffroy de Clippel[†] Zhuorong Mao[‡] Kareen Rozen[§]

September 2026

Abstract

This paper studies whether AI-augmented choice overcomes cognitive constraints, or whether these constraints continue to shape how individuals interact with AI and hamper their ability to reap its benefits. Our experimental design varies AI access in a series of tasks based on [Choi, Laibson, and Madrian \(2010\)](#)’s classic portfolio-choice problem. Though AI access improves performance and learning, performance remains well below the near-perfect outcomes AI can achieve with a simple prompt. Leveraging rich conversational transcripts with AI, we find that participants often use less efficient modes of engagement with AI. We trace their performance shortfall to two additional human bottlenecks: information provision and information extraction. Participants often provide incomplete descriptions of their tasks and only partially synthesize the information AI provides. To quantify the information-extraction bottleneck, we benchmark participant performance against the performance AI could have achieved using the same resources (e.g., the conversational transcript). Overall we find that in AI-augmented choice the cognitive burden shifts from *solving* the problem to *conveying* it to AI and *synthesizing* the response.

Keywords: AI, bounded rationality, cognitive costs, complexity, delegation, investment, finance

JEL codes: C91, D91, G11

*We are grateful to Chris Jeong for excellent research assistance. This study received approval from Brown University IRB. We thank Brown University’s Center for Computing and Visualization for facilitating access to Microsoft Azure. On AI use: a chatbox was embedded into some screens of the experimental interface for subjects’ use; AI was used for textual analysis of user prompts as well as for simulations of AI performance in the investment problem; and AI was consulted for coding assistance and wording improvements.

[†]de Clippel: Economics Department, Brown University, declippel@brown.edu.

[‡]Mao: Economics Department, Brown University, zhuorong_mao@brown.edu.

[§]Rozen: Economics Department, Brown University, kareen_rozen@brown.edu.

1 Introduction

A central finding in behavioral economics is that complexity and cognitive constraints can impair the quality of economic decisions. The widespread availability of artificial intelligence (AI) has, however, altered the decision landscape: individuals can use AI tools to help process information and generate recommendations at little cost. But whether access to AI overcomes human limitations in decision-making is not obvious. When people engage freely with AI, part of their cognitive burden shifts from *solving* the problem to *conveying* the problem to AI and *synthesizing* its response. This leaves open the possibility that cognitive limitations also shape how individuals interact with AI, hindering their ability to reap its benefits. We investigate this possibility in the present paper.

We conduct a preregistered and tightly-controlled laboratory experiment in which participants make a series of complex economic decisions, systematically varying access to AI. Building on a realistic task described below, our design allows participants to engage freely with AI for assistance, rather than accept or reject a predetermined recommendation. We leverage task performance, conversation transcripts with AI, and responses in complementary elicitation tasks to study several key questions. First, how do people use AI to solve complex problems? Second, does AI improve decision quality, learning and meta-cognition of performance? Third, do cognitive constraints hamper performance in AI-assisted environments, and if so, which human bottlenecks play a role?

In designing the experiment, we sought an economically relevant real-world task that is challenging, permits objective measurement of decision quality, and naturally lends itself to diverse forms of AI assistance. For this, we build on the classic investment task of [Choi, Laibson, and Madrian \(2010\)](#): portfolio allocation among S&P 500 index funds. These funds track the same index and thus have identical gross returns, but vary in their annual and one-time costs (such as ‘front loads’, which are upfront reductions in the principal invested). Over a short investment horizon like a year, one-time costs dominate any intertemporal considerations, and agents should invest only in the fund that minimizes these costs.

The optimal solution to this investment problem is thus unambiguous; nonetheless, it is elusive in practice. Motivated by empirical evidence of overinvestment in high-fee index funds, [Choi, Laibson, and Madrian \(2010\)](#) find that their experimental participants (college students recruited at Harvard, Wharton MBA students and Harvard staff members) perform poorly, consistently investing in suboptimal funds. Indeed, there are many hurdles to identifying and implementing the optimal solution. For instance, one must understand the fee structure of index funds, comprehend financial jargon and summary statistics, overlook extraneous information (e.g., returns since inception) and pinpoint the relevant information in dense and lengthy fund prospectuses.

AI may alleviate these obstacles in multiple ways. Subjects could use AI as a substitute for their own cognition by *requesting AI’s recommendation* of the best portfolio; they could complement their

cognition by *seeking guidance* and improving their conceptual understanding of the task; or they could use it to reduce their cognitive costs by simply asking it to *retrieve information* from summary prospectuses.

Yet cognitive constraints may continue to matter under all three forms of engagement. When individuals seek guidance, they must exert effort to synthesize AI’s responses. When they ask AI to retrieve information, the usefulness of AI’s response depends on whether they identified the relevant information to request and can incorporate it correctly. The role of cognitive constraints may be less obvious for subjects who simply ask AI to tell them what to choose. In principle, these subjects could achieve very high performance at very low cost. But even delegating the problem to AI may not eliminate the need for cognition: AI can solve only the problem that the user conveys. The user must recognize what information in the task is important and make efforts to specify it. Indeed, as we establish in simulations, the AI tool available in our experiment (o4-mini) performs nearly perfectly in this task with a simple prompt providing key features of the decision environment (e.g., the length of the investment horizon), but performs significantly worse when the decision environment is incompletely specified.

Our first finding is that access to AI substantially improves decision quality. Participants without AI access on the initial investment task spend considerable time and engage with the summary prospectuses, yet their choices are statistically indistinguishable from those generated by an uninformed strategy, suggesting that they gain little understanding of how to optimize their returns. By contrast, participants with AI access perform significantly better, and we find some evidence that these benefits extend beyond the immediate task, with subsequent improvements in performance.

Decision quality with AI access nevertheless remains far below what AI itself can achieve. Participants perform substantially worse, despite using the same AI model that performs nearly perfectly in simulations. We find that this gap between AI capability and AI-assisted human performance is systematic, raising a critical empirical question: what prevents people from realizing the full potential of a capable AI assistant?

We identify three main human bottlenecks that contribute to this shortfall.

A first bottleneck is that participants may *choose an inefficient mode of interaction* with the AI. We find that nearly all AI interactions in our experiment fall into one or more of our pre-registered categories: recommendation-seeking, guidance-seeking, and information-seeking. Although a simple recommendation-seeking prompt can ensure near-optimal outcomes, many participants engage in more cognitively-demanding conversations that, on their own, are less informative for reaching an optimal choice. Such behavior may be sensible if participants enjoy learning about the task, view the AI as a helpful tool rather than a decision-maker, are unaware of the best way to interact with AI, or are simply skeptical of its recommendations. Consistent with this last possibility, we find that subjects uniformly underestimate AI’s accuracy. Nonetheless, average performance remains

suboptimal across all modes of interaction, including among subjects who seek recommendations. This suggests additional sources of inefficiency, beyond the chosen mode of interaction, contribute to the shortfall.

A second bottleneck concerns *information provision* to AI. Even when seeking recommendations, participants frequently fail to provide AI with all the information needed to generate accurate recommendations, such as the investment horizon or the feasible set of available funds. Such omissions substantially reduce their performance. This finding is consistent with AI’s inferior performance in our simulations for an incompletely-specified decision environment. In such cases, the AI may implicitly fill in missing information with incorrect assumptions, such as whether the investment horizon is short or long, or which funds listed on the summary prospectuses are actually available.

In contrast to the prior two bottlenecks, the third arises after the AI responds and concerns *information extraction*. Even conditional on the contents of the conversation transcript, summary prospectuses, or task description, participants may fail to translate that information into decisions as effectively as a capable AI equipped with that same information. We quantify the informativeness of subjects’ available resources through a simulation exercise. For this, we deploy the same, highly accurate AI model from our experiment to estimate the performance that it, rather than a human, would achieve when provided different subsets of our subjects’ informational resources. We use these AI-derived measures of informativeness as a yardstick to assess subjects’ performance, computing for each subject their *AI-benchmarked information utilization*. This analysis reveals that information extraction is a significant bottleneck. Participants appear unable to extract any information from the prospectuses beyond what is contained in the transcript, and extract less than half the information available in the transcript itself. Performance is thus adversely impacted both because subjects often incompletely specify the problem, and because they only partially synthesize the information that AI provides.

Beyond performance, the process data are consistent with boundedly rational subjects shifting their effort and attention. Participants engage substantially with AI. In sum, they spend significantly more time on the task itself while sharply reducing their engagement with primary information sources (the summary prospectuses).

Though participants report higher confidence in the optimality of their decisions when provided with AI access, the evidence on their ability to assess decision quality is more subtle. Our final two tasks provide complementary evidence on how they assess their own decisions and those of AI. One task elicits participants’ willingness to pay for an optimal intervention: receiving the optimal profit instead of the profit achieved by their portfolio allocation in the preceding investment task. For participants who chose that portfolio *without* AI access, willingness to pay declines with portfolio performance. By contrast, this expected relationship disappears for participants who chose the portfolio when AI access was available. The other task elicits beliefs about AI’s accuracy in

identifying the optimal fund in a previously-conducted simulation, using a prompt that provides all the necessary details. Participants greatly underestimate AI’s accuracy, although the vast majority believe that AI is more likely than not to identify the optimal fund. This suggests that cognitive uncertainty about how to solve the investment task extends to the ability to evaluate AI’s solutions. On balance, participants correctly expect AI to improve their performance, but seem better able to gauge the quality of their decisions when working independently.

Related literature

Our paper contributes to three main literatures.

Cognitive constraints in individual decision making. First, we contribute to a literature in behavioral and information economics that investigates the factors that make individual decision-making difficult. The importance of information acquisition has been recognized by economists since [Stigler \(1961\)](#), and models of rational inattention further emphasize that the processing of information is itself costly; see [Sims \(2003\)](#) and [Caplin and Dean \(2015\)](#), among others. Their central insight is that access to information need not translate into knowledge. Decision making can also be difficult due to the overarching problem of choosing *how* to approach the problem in the first place: which aspects to attend to, what information to seek, and what procedure to use to reach a decision. [Simon \(1976\)](#) suggests that realistic decision making is often procedural, arriving at a merely satisfactory solution. Concrete economic models in this direction include [Salant \(2011\)](#), who analyzes the choice of decision procedures under bounded rationality, and [Gabaix \(2014\)](#), who models how individuals create sparser and more tractable representations of a problem.¹

The investment task of [Choi, Laibson, and Madrian \(2010\)](#), on which our experiment builds, illustrates several of these challenges. To maximize payment, a participant must identify the optimal strategy, understand cost structures, distinguish relevant from irrelevant information, retrieve key details from lengthy prospectuses and correctly compute and minimize costs. Broadly, the task is complex in the sense studied by a recent and growing literature on cognitive costs and complexity in economic decision-making; see surveys by [Enke \(2024\)](#) and [Oprea \(2026\)](#). Its difficulty is evidenced by poor performance even among highly educated participants. [Choi, Laibson, and Madrian \(2010\)](#), and related work by [Beshears, Choi, Laibson, and Madrian \(2011\)](#), explore whether informational interventions such as simplified disclosures and financial education improve decision quality, finding that any improvements are modest at best.

Generative AI has the potential to fundamentally alter the decision-making process by helping users reason through complex decisions and facilitating the retrieval and synthesis of information. At one extreme, users can delegate a problem to AI and merely decide whether to accept its recom-

¹See [de Clippel and Rozen \(2024\)](#) for a survey of bounded rationality in choice theory

mendation; at the other, they can retain ownership of the decision-making process while consulting AI for guidance or using it to lower information retrieval costs. Nearly all our participants' interactions with AI fall into these categories, and access to AI does substantially improve their decision quality. At the same time, their performance remains well below the near-optimal performance achieved by the AI using a simple delegation prompt.

This gap highlights our contribution to the literature on behavioral and information economics. Once the task is appropriately specified, the AI can perform much of the cognitive work needed to optimally solve it. The participant, however, is responsible for determining how to engage with the AI, specifying all relevant inputs, and interpreting AI's output. We show that these residual cognitive demands substantially limit the value of the technology. Participants frequently adopt interaction strategies that are more cognitively demanding than necessary; they omit information needed to obtain accurate recommendations; and they manage to extract only a fraction of the information contained in the AI's responses. Hence generative AI does not fully eliminate the cognitive costs of decision-making; instead, it shifts them elsewhere.

Machine learning and recommendations. Our paper also relates to a literature applying machine-learning methods for prediction and recommendations. [Kleinberg, Ludwig, Mullainathan, and Obermeyer \(2015\)](#) point to a wide range of consequential settings where accurate predictions are of the essence: judges assess the risks of releasing defendants, physicians assess the likely outcomes of treatments, lenders assess the likelihood of repayment, etc. [Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan \(2018\)](#) demonstrate the potential value of machine predictions and the challenges inherent in evaluating whether they improve upon human decisions.

In many of the applications studied, the recommendation algorithms are used to assist rather than replace human decision makers, with mixed results. For example, [Gruber, Handel, Kina, and Kolstad \(2021\)](#) document improvements in the performance of skilled health-insurance experts, while [Agarwal, Moehring, Rajpurkar, and Salz \(2023\)](#) find zero average effect for radiologists, a result they attribute to underweighting of AI's signal and neglecting the correlation with one's own information. [Angelova, Dobbie, and Yang \(2026\)](#) similarly show that judicial overrides of algorithmic recommendations usually reduce decision accuracy, with only a small minority of judges applying their private information helpfully. [Caplin, Deming, Li, Martin, Marx, Weidmann, and Ye \(2025\)](#) show the benefit from algorithmic recommendations in an age-prediction task depends on how well-calibrated a subject is regarding their own ability. Other work documents additional behavioral responses that impact the benefits of algorithmic advice. [Dietvorst, Simmons, and Massey \(2015\)](#) see a reluctance to rely on algorithms after a mistaken recommendation; and [Agarwal, Moehring, and Wolitzky \(2026\)](#) observe both under-responsiveness to recommendations (due to overconfidence in the precision of personal information) and crowding-out of human effort (especially when AI is confident in its recommendation). Motivated by such evidence, [McLaughlin and](#)

[Spiess \(2024\)](#), [Agarwal et al. \(2026\)](#) and [Hoong and Dreyfuss \(2026\)](#) study the design of recommendation systems while taking into account behavioral responses.

In the above works, inputs to the prediction algorithm, and the algorithm itself, are pre-specified by the institutional or technological setting. The AI either directly observes all relevant information or obtains it through a structured intake process. This differs in multiple respects from consumer-facing large language models (LLMs) as we study here. By contrast to institution-specific algorithms, these LLMs are widely available to individuals, and are multi-purpose tools that can be applied to nearly any setting. They allow for an organic form of interaction, with the human and AI engaging in an open-ended exchange. The potential scope of assistance in daily choices is thus broader: the human can ask the AI to provide facts, explain how to approach a problem, perform cognitive operations, or make a recommendation. The cost of this flexibility is that it introduces new sources of inefficiency. The human must define their problem and provide inputs to enable AI to perform the cognitive work; and afterwards, the human must interpret and act on AI's response. Our paper studies these aspects of general-purpose AI interactions and their implications for decision quality.

Human-generative AI collaboration. Finally, our paper contributes to the emerging literature on human-generative AI collaboration in a broad range of settings. [Noy and Zhang \(2023\)](#) show that ChatGPT improves speed and quality in professional writing tasks; [Brynjolfsson, Li, and Raymond \(2025\)](#) find productivity gains when workers in customer support are offered recommended phrases by a generative-AI conversational assistant that monitors the conversation; and [Choi and Schwarcz \(2025\)](#) examine pedagogical use of GPT assistance in legal exams, documenting large gains for structured questions but not for complex writing tasks (though they do not systematically analyze students' prompting behavior, they conjecture this is due to how students use the technology). These studies establish that LLMs can substantially augment cognition, while also highlighting that the gains depend on the task. Studying an array of tasks drawn across academic fields, [Vafa, Ram-bachan, and Mullainathan \(2024\)](#) also show that people can deploy AI for the wrong tasks due to miscalibrated beliefs about which tasks AI performs well. Rather than which tasks people choose to delegate, we study how people interact with AI in a fixed task.

More relevantly, our paper joins a body of work that has begun to study how the use of generative AI shapes economic decision making. [Chen, Filiz-Ozbay, and Ozbay \(2026\)](#) examine whether access to an LLM recommendation impacts how people value risky, uncertain and complex bets; the optimal choice is subjective in their setting, which is based on [de Clippel, Moscariello, Ortoleva, and Rozen \(2025\)](#). Like [Brynjolfsson, Li, and Raymond \(2025\)](#), and similar to the machine-learning literature, the recommendation in [Chen, Filiz-Ozbay, and Ozbay \(2026\)](#) is generated without any

user prompting. Other works study the act of prompting itself.² Choukhmane, de Silva, Lin, and Akuzawa (2026) ask a representative sample to write a prompt describing their financial situation and requesting spending and investment advice from an LLM, which the authors apply in GPT-5.2 to simulate outcomes in a parameterized life-cycle model; they find differences in LLM recommendations and simulated wealth accumulation based on several personal characteristics, and especially gender (they attribute this mostly to gender differences in the prompts themselves, but partially to gender information passed alongside even identical prompts). Fedyk, Kakhbod, Li, and Malmendier (2026) compare stock, bond and cash ratings simulated by AI to those elicited from human participants in a large-scale survey; they find that AI’s default ratings often resemble those of young, high-income investors, but that providing demographic information in the simulation’s prompt improves the alignment with ratings elicited from the corresponding demographic groups. Imas, Lee, and Misra (2026) experimentally construct an agentic market where humans are represented by buyer- and seller-side AI agents, and show that people differ greatly in their ability to instruct these agents effectively, generating what they call a “specification hazard.” Studying prompting and parsing of information in a field experiment with entrepreneurs in Otis, Clarke, Delecourt, Holtz, and Koning (2026) provide access to a GPT-4-based business assistant which, when asked a question, offers three to five suggestions that vary in their helpfulness. They find no differences in the topics (discounting, advertising, etc.) raised between previously high-performing (above-median profit in the pre-treatment phase) and previously low-performing (below-median pre-treatment profit) entrepreneurs; and correspondingly, no meaningful differences in the AI assistant’s responses between these groups. But previously high-performing (low-performing) entrepreneurs are better (worse) at discerning between helpful and unhelpful recommendations; this leads them to have higher (lower) post-treatment profits, and results in zero average effect on profits overall.

In our experiment, performance is dependent not only on prompting, but also on the ability to parse AI’s response. Importantly, our task has an objective optimal solution, which allows us to gauge the efficiency of any choice relative to the optimum, a feature not shared by some of the works above. This allows us to not only evaluate the performance of participants, but also to evaluate AI performance under idealized and imperfect prompting conditions, as well as when provided with the same conversational transcripts and other sources available to our participants. These measures permit us to compute AI-benchmarked information utilization, or how much information participants manage to extract and act on, relative to what a capable AI could have (counterfactually) achieved under the same conditions. We thus study both initial problem specification to AI and downstream information extraction from AI (and other sources) within the same economic setting,

²Yet other work studies how decisions are impacted by the style of AI’s own communication. Conlon and Schwardmann (2026) find that AI advice depolarizes choices despite being sycophantic, and that increasing sycophancy weakens depolarization. Dreyfuss and Raux (2026) find that the more anthropomorphically-packaged an AI tool is, the more humans project their own abilities onto it.

benchmarking both against AI’s performance.

2 Experimental design

Our experimental design varies access to AI during a portfolio choice task, with participants freely deciding how to interact with the AI, if at all. It investigates how different modes of engagement, ranging from guidance seeking to recommendation requests, impact decision quality; investigates the extent to which cognitive constraints bound the benefits of AI; and studies whether these interactions produce durable knowledge. The experimental design was preregistered on AsPredicted.org.³

We first introduce the basic building block of our experiment before describing the full series of tasks and key treatment variations. Screenshots of the instructions and interface are provided in Online Appendix A.

Core investment task. As in Choi, Laibson, and Madrian (2010), individuals allocate a fixed amount across S&P 500 index funds. In the investment task, participants allocate \$12,000 across four S&P 500 index funds for a fixed time horizon (June 4, 2024 to May 6, 2025). They are free to choose any ‘valid’ allocation over these funds, meaning that (i) the total investment is \$12,000 and (ii) if allocating a non-zero amount to a fund, then that amount must satisfy the fund’s minimum opening requirement. Each menu of four S&P 500 index funds contains two higher-cost funds (5.00%–6.28%) and two lower-cost funds (2.04%–3.16%) over the experiment’s 11-month investment horizon.⁴ The payoff associated with an investment task in our experiment is 5% of the net return of the chosen portfolio over the time horizon. Participants are informed that returns are guaranteed to be nonnegative over this time period.

Focusing on S&P 500 index funds allows for a clear optimal benchmark: since all funds track the same index, their gross returns are nearly identical. Thus, the optimal choice is the fund with the lowest cost, which delivers the highest expected net return. Each participant completes three investment tasks (Q1, Q2, and Q3). In each task, the four-fund menu is drawn from a common pool of twelve S&P 500 index funds, listed in Appendix A. The composition of each four-fund menu is randomized at the individual level without replacement.⁵

Informational conditions and AI access. Participants complete the three investment tasks (Q1, Q2, and Q3) under varying informational conditions.

³The protocol can be found at <https://aspredicted.org/9ti2q3.pdf>.

⁴Costs listed here are approximated to a first order by the load plus the annual expense ratio multiplied by 11/12. Investment returns used in the analysis are calculated by applying each fund’s growth rate, net of expense ratio, to the amount invested after deducting the load.

⁵Menus are drawn without replacement subject to the constraints that each menu contains two high-cost and two low-cost funds, and that the funds are offered by different companies.

You are asked to allocate \$12,000 among the S&P 500 index funds listed in the table below. Recall that we will calculate your payoff as if your investment is made at 3pm on June 4, 2024 and sold at 3pm on May 6th, 2025 (with dividends reinvested along the way). You can click on a fund's name to view its summary prospectus.

Enter your decision in the last column of the table at the bottom of this page. You may invest in as many funds as you'd like (that is, you can invest in just one, a pair of funds, three funds, or all four). If you do invest in a fund, then you must satisfy its minimum opening requirement.

Mutual Fund	Ticker	Minimum Opening Requirement (if all buying)	Your Allocation
DWS S&P 500 Index Fund Class A	SXPAX	\$1,000	0
Victory S&P 500 Index Fund Class A	MUXAX	\$2,500	0
Rydex S&P 500 Fund Class C	RYSYX	\$2,500	0
Invesco S&P 500 Index Fund Class A	SPIAX	\$1,000	0
Total:			\$0

[Validate and Save](#)

Figure 1: Baseline decision interface (without AI chatbox).

- *Baseline: no AI access.* Participants only have access to the contemporaneous summary prospectuses available at the investment start date for the four available funds. As seen in Figure 1, the decision screen includes all relevant information to the task: the time horizon, incentive scheme, the four available S&P 500 index funds (the display order of the funds is randomized), and any minimum opening requirements. By clicking on a fund name, the fund's summary prospectus opens in a modal overlay (i.e., a window which blocks the interface until closed), allowing us to measure engagement with the prospectus by the duration of time the document is open. Subjects enter amounts for each fund, with the total amount allocated updated and displayed in real time. The interface validates the allocation (requiring any invalid allocation to be corrected) before the subject can proceed to the next task.
- *Treatment: AI chatbot.* The decision screen contains all elements above, *plus* a custom chat interface connected to OpenAI's o4-mini model (a common versions of OpenAI's ChatGPT, which is recommended for coding, math and science uses). The chatbox is embedded between the description of the task and the list of available mutual funds. Participants are informed that the o4-mini model version has a May 2024 knowledge-cutoff date, so the AI does not have access to realized fund performance over the investment period. Participants may freely interact with the chatbot through unrestricted text prompts. All four summary prospectuses are automatically uploaded to o4-mini, along with a system prompt (provided in Online Appendix A). The system prompt only communicates that (i) this is an experiment on financial decision-making among S&P 500 funds, (ii) the subject may be paid in proportion to their hypothetical net return, and (iii) the prospectus uploads are automatically provided by the experimenter, so the AI does not treat their presence as a signal to elaborate on their content unless the subject explicitly requests it. Participants are shown this system prompt in the instructions, and it is displayed as the first prompt in the chatbox.⁶ For realism, this system prompt *does not* provide AI with the investment time horizon or which specific share

⁶Beyond this initial prompt, the chatbox displays all prompts and responses between the subject and AI for the current investment task, and the subject can scroll through these.

classes are available from among the uploaded prospectuses. This reflects an important feature of natural settings where users must actively⁷ communicate relevant features of their problem to AI themselves. All correspondence between the participant and AI is recorded as part of the dataset.

Informational conditions in the three investment tasks (Q1-Q3) vary across treatment arms. To ensure tight control over these conditions, the experiment was implemented in-person, with internet access at computer terminals shut down (aside from any domain needed to run the survey itself).⁸ The design has four treatment arms that vary the timing of AI access across the three investment tasks (each subject is assigned to one treatment). These are shown in Table 1 below. In the T1 treatments (T1-, T1+), participants have access to AI in Q1 but not Q2; in the T2 treatments (T2-, T2+) this ordering is reversed. The T1 and T2 groups are subdivided into “plus” treatments (T1+, T2+) and “minus” treatments (T1-, T2-): participants in the plus treatments have access to AI in Q3, while those in the minus treatments do not. Importantly, there is no opportunity for feedback during the experiment other than the opportunity to interact with AI.

Treatment	Q1	CU	Q2	Q3	WTP	Belief
T1-	AI	✓	No AI	No AI	✓	✓
T1+	AI	✓	No AI	AI	✓	✓
T2-	No AI	✓	AI	No AI	✓	✓
T2+	No AI	✓	AI	AI	✓	✓

Table 1: Timeline of tasks and AI access across treatment arms.

Our design sets up clean tests for our key questions in the next section, namely: how AI access impacts initial decision quality (Section 3.1); how experience, both with and without AI, impacts performance (Section 3.2); what modes of interaction do subjects use when engaging with AI (Section 3.3); and what, if any, are the human bottlenecks that may limit the ability to fully reap AI’s benefits (Section 3.4). Additional elements in the design allow us to study important complementary questions.

CU. In the task immediately following Q1 (denoted CU in Table 1), we measure participant’s cognitive uncertainty about their optimality of their Q1 portfolio decision using a (non-incentivized) subjective confidence question adapted from [Enke and Graeber \(2023\)](#). This allows us to assess the impact of AI on perceived decision quality.

⁷The ability to copy-paste text from the instructions and decision screens was disabled to ensure that the problem description subjects pass to the chatbot is actively chosen.

⁸The survey is run on Qualtrics, and the chatbox accesses o4-mini by passing through a university server.

We also study whether AI assistance impacts the valuation of optimal interventions, and how it shapes the link between valuation and the intervention’s actual benefit. The variation in AI access in Q3 across the plus and minus treatments is relevant to the analysis of the WTP task we now describe.

Willingness to pay (WTP) for optimal intervention. After completing Q3, participants complete a separate payoff-relevant task. Specifically, they choose whether the payment for the WTP task will be based on the Q3 portfolio they just selected or, for a fee, based on the portfolio that maximizes profit in Q3. We ask each participant which option they prefer for a series of fees varying from \$1 to \$25 (in steps of \$2). Participants in the ‘plus’ treatments had AI access when selecting their Q3 portfolio, while those in the ‘minus’ treatments did not. Differences in WTP across these groups reflect how AI assistance shapes participants’ valuation of an optimal intervention. Moreover, WTP for the intervention can be compared with how much a participant *should* have paid, allowing us to evaluate how AI assistance impacts the link between performance and self-assessment. The payoff for this question is determined by selecting one row uniformly at random and implementing the subject’s choice in that row.

Beliefs about AI accuracy. The last main task, denoted Belief in Table 1, examines participants’ trust in AI’s judgment. Participants are asked to report their best guess of o4-mini’s accuracy in 1,000 simulations in which o4-mini was prompted to select the net-return–maximizing fund from a two-fund menu (drawn randomly each time from the twelve funds encountered in the experiment) over the same investment horizon used in Q1-Q3. Participants’ beliefs about AI’s percentage of correct responses were elicited using a binarized scoring rule (BSR) with a \$10 prize.

We gather rich choice-process data. In addition to the decision data, we have a record of all messages (prompts and replies) in the AI chatbox. We record the total time subjects spend on each question, which allows us to compare decision efficiency between treatments and to examine the relationship between time investment and decision quality. Furthermore, we record the time each prospectus is open in Q1-Q3. We interpret this measure as engagement with the prospectus—that is, effort devoted to directly retrieving and processing information from the prospectuses—as the modal overlay prevents returning to the decision interface until the prospectus is closed. A significant reduction in prospectus viewing time in the presence of AI access would suggest that AI substitutes for, rather than complements, subjects’ information acquisition and processing effort.

After the main survey, subjects complete an optional exit survey. Subjects self-report their level of prior financial knowledge and prior experience with LLMs, allowing us to investigate whether the effect of AI on decision quality is mediated by these characteristics. We also collect information on subjects’ motivations for using AI during the task (e.g., delegating the task, identifying an approach,

or extracting information) and how helpful they found the AI assistance. Finally, we collect basic demographic information, including gender, concentration, and number of semesters enrolled as Brown University undergraduates.

Payment. One of the five incentivized questions (Q1, Q2, Q3, the MPL task, or the Beliefs about AI accuracy task) is randomly drawn to determine the subject’s variable earnings (received in addition the fixed \$10 show-up payment). Subjects receive all payments in cash at the end of the session.

Implementation details. The experiment was conducted in person at the Brown University Social Science Experimental Laboratory, with the survey implemented in Qualtrics. The experimenter read aloud a brief handout (Online Appendix A) that describes general procedures for the session (e.g., putting away electronic devices, how payment is received). The experimenter then provides a passcode participants need to begin the Qualtrics survey. Each question in the survey was preceded by detailed instructions.

In each session, terminals were assigned to one of the four treatments through an embedded variable in the Qualtrics survey link, with treatments rotating across terminals in sequence (e.g., T1-, T1+, T2-, T2+, T1-, T2+, T2-, T2+, ...). To ensure a near-even distribution of subjects across treatments, the first terminal used in each session was assigned the treatment that follows the last-used treatment in the prior session.⁹ Internet access in each computer terminal was blocked, with the exception of two domains needed to run the survey (Qualtrics and Brown, with the latter used to route traffic back and forth between the terminal and Microsoft Azure, which hosted the 2025-04-16 version of o4-mini used for the interface).

Following the preregistration, sessions were conducted until there were at least 40 participants per treatment. This process yielded 16 sessions between November 2025 and February 2026. Subjects could participate in only one session. The dataset contains 162 participants, with 41, 41, 40, and 40 in T1-, T1+, T2-, and T2+, respectively. This yields 82 participants in the T1 group, 80 in the T2 group, and 81 in each of the plus and minus groups.

3 Main Results

3.1 The Impact of AI Access on Initial Decision Quality

The most direct way to assess the potential value of AI assistance is to compare decision quality across treated groups (T1 vs. T2) in the first question of our survey (Q1). Focusing on Q1 provides

⁹Three subjects who encountered technical issues with the chatbot (in earlier sessions) were excluded and replaced with three subjects in the final session; these replacements were assigned the same treatment the replaced subjects took, to maintain balanced sample sizes across treatments.

a clean between-subject comparison, as it captures participants' initial responses prior to any exposure, learning, or adaptation in subsequent questions. Later questions can incorporate experience effects that we study explicitly below, but Q1 offers the simplest benchmark (analogous to a single between-subject, treatment–control experiment) for measuring the immediate impact of AI access.

Result 1 shows that the allocation problem is challenging: on balance, participants fail to outperform uninformed decisions (in the sense formalized below) despite spending substantial time on the task.¹⁰ Result 3 shows that participants benefit significantly from AI assistance, yet remain far from the near-optimal performance attainable with a simple, fully specified delegation prompt (see Result 2).

A (stochastic) *allocation rule* assigns to each menu S of four index funds a probability distribution over investment shares across the funds in S . The *optimal rule*, for instance, assigns to S the Dirac measure on the cost-minimizing fund. An allocation rule is *uninformed* if, for every menu S , the associated distribution is symmetric across funds, that is, uninformed rules ignore fund characteristics. Examples include equal splitting, uniform random allocations, or investing entirely in a randomly selected fund.

Result 1. *Participants without AI assistance in Q1 exert substantial effort, but fail to outperform the uninformed benchmark.*

Supporting Evidence. These participants spend, on average, more than nine minutes on Q1. This includes an average of around 106 seconds browsing each prospectus opened, with 88.1% viewing all four prospectuses.

Despite this effort, performance is low. We provide two pieces of evidence. First, all uninformed rules imply equal expected shares. Averaging across participants without AI access in Q1, the share invested in the lowest-cost fund is 27.58%, with a share of 50.19% invested in the two lower-cost funds. These are statistically indistinguishable from 25% and 50% respectively (p-values: 0.27 and 0.95).

Going a step further, we estimate a simple, intuitive mixture model of uninformed behavior. Let λ_k be the share of participants attending to $k \in \{1, 2, 3, 4\}$ funds. Conditional on k , the consideration set of funds is drawn uniformly from among all size- k subsets of S , and allocations follow a symmetric Dirichlet distribution over the consideration set, with parameters $\alpha \mathbf{1}_k$. This parametrization nests uniform random allocation ($\alpha = 1$) and converges to equal splitting as $\alpha \rightarrow \infty$.¹¹ Under this specification, participants invest with probability one in every fund they consider. Accordingly, the MLE of λ_k has a natural interpretation: it represents the proportion of participants who invest in exactly k funds. Maximum likelihood estimates yield $\lambda_1 = 8.75\%$, $\lambda_2 = 13.75\%$, $\lambda_3 = 31.25\%$,

¹⁰A similar pattern appears in Choi et al.: college students in the prospectus-only condition incurred average portfolio fees of \$431 on a \$10,000 investment, only slightly below the \$443 implied by random choice.

¹¹We impose a common α across k for parsimony.

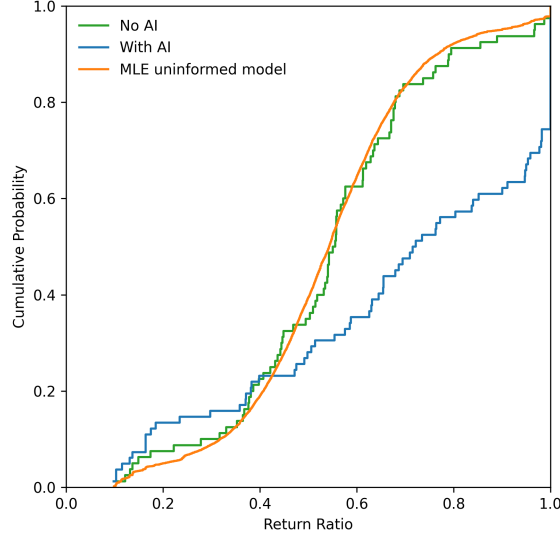


Figure 2: Q1 Performance CDFs: With AI access vs. without vs. MLE uninformed model.

$\lambda_4 = 46.25\%$, and $\alpha \approx 4$. Figure 2 shows that the empirical and simulated performance CDFs are nearly identical, showing performance in Q1 without AI is indistinguishable from uninformed choice despite the substantial effort exerted. ■

We next show that near-optimal allocations are, in principle, achievable when interacting with o4-mini. *Delegation* consists of asking the AI to choose the allocation and following its recommendation. *Full-information delegation* additionally provides the exact investment horizon and set of available funds.¹²

Result 2. *Full-information delegation yields near-perfect performance in simulations.*

Supporting Evidence. We ran 1,100 simulation trials (ten for each valid four-fund menu that can be constructed from the 12 funds used in the experiment). When provided with complete information, o4-mini achieves an average performance of 99.45%. It selects the optimal allocation, investing all \$12,000 in the cost-minimizing fund, in 95.45% of trials. ■

Result 3. *Participants' performance significantly improves when AI is available in Q1, though it still remains far from optimal.*

Supporting Evidence. We begin by noting that nearly all (90.2% of) participants with AI access in Q1 do indeed converse with AI.¹³ The average performance with AI access in Q1 is 68.5%,

¹²Recall that prospectuses are automatically uploaded to the AI, but specifying the relevant share class remains necessary as prospectuses typically cover multiple class shares from the same firm.

¹³AI takeup rates in Q2 and Q3 are similarly high: 93.8%, and 90.1%, respectively.

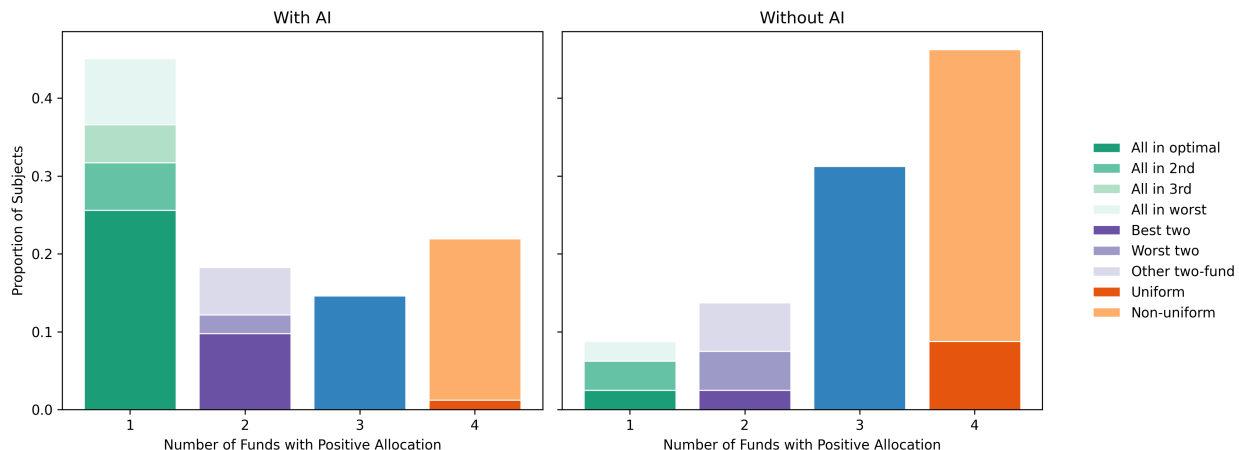


Figure 3: Overview of Q1 allocation decisions per treatment.

which is significantly larger than the 54.8% average performance in Q1 when no access is provided ($p = 0.001$). Moreover, with AI access the average share in the best fund increases to 43.92%, with an average share of 65.23% invested in the two best funds. Both significantly exceed the uninformed benchmarks (25% and 50%, respectively) at the 1% level. Overall, the distribution of performance with AI effectively dominates the distribution of performance without it, as seen in Figure 2.

Yet, even with AI, performance in Q1 falls well below the near-perfect distribution of performance achieved in the full-information delegation simulations. This can be seen at once by comparing the above performance statistics with those in Result 2. ■

Overall, there is a stark contrast between decisions made with and without AI assistance, as seen in Figure 3. When decisions are unaided (the right panel) there is a tendency to invest in multiple funds, which may reflect a misguided belief that diversification is optimal or an inability to discern between funds. There is a near reversal in behavior with AI (the left panel): even when unsuccessful at identifying the best fund, participants are far more likely to invest in only one fund.

3.2 Effects of Experience on Performance

Beyond the impact of AI assistance on initial decisions, we now study the effects of experience. The results below reveal a sharp asymmetry. Prior unaided experience does not improve subsequent performance with AI assistance. On the other hand, prior AI-assisted experience substantially improves subsequent decisions, both when AI remains available and when it does not. These findings suggest that interacting with AI teaches participants something economically meaningful about the allocation problem itself and may help them use AI more effectively in future interactions.

Recall from Result 1 that participants in Q1 without AI devote substantial time to the task (despite performing poorly). One may wonder whether this experience helps them better leverage AI assistance in subsequent choices. To examine this question, we compare AI-assisted performance between participants with prior unaided experience (Q2 choices in T2) and participants encountering the task with AI assistance for the first time (Q1 choices in T1). We find no discernible effect of prior unaided experience on subsequent AI-assisted performance.

Result 4. *When AI is available, there is no difference in performance between participants with prior unaided experience and those with no prior experience.*

Supporting Evidence. The share invested in the cost-minimizing fund (top two funds) is 49.01% (resp., 68.36%) for participants with prior unaided experience (Q2 in T2), compared to 43.92% (resp., 65.23%) for participants interacting with AI for the first time (Q1 in T1). The differences are small and statistically insignificant (p-values 0.4164 and 0.5922). ■

We next examine the reverse channel: whether making an AI-assisted choice improves subsequent unaided decisions. Specifically, we compare AI-free performance among participants who previously completed a similar task with AI assistance (Q2 choices in T1) to performance among participants facing the task without any prior exposure (Q1 choices in T2).

Result 5. *In the absence of AI, participants with prior AI-assisted experience outperform those with no prior experience.*

Supporting Evidence. The share invested in the cost-minimizing fund (top two funds) is 40.19% (resp., 60.77%) among participants making their unaided choice in Q2, compared to 27.58% (resp., 50.19%) among participants making their unaided choice in Q1. These differences are both economically meaningful and statistically significant (p-values 0.0098 and 0.0349). ■

Another, related way to assess the impact of AI exposure is to compare combined Q1 and Q2 performance in T1 and T2. Participants in both treatments answer two allocation questions, one with AI access and one without. The only difference is sequencing: participants in T1 interact with AI before making an unaided choice, whereas participants in T2 encounter AI only after first attempting the task on their own.

Result 6. *Combined Q1 and Q2 performance is higher when AI assistance is provided first (T1) rather than second (T2), though this effect is limited to the top tail of performers. In particular, participants who achieve near-optimal allocations in both questions belong to T1.*

Supporting Evidence. Figure 4 shows that, when performance in Questions 1 and 2 is combined, treatment group T1 exhibit a larger share of high-performing subjects, consistent with persistence of the AI effect beyond the initial task. In line with this pattern, Figure 4 also shows that subjects

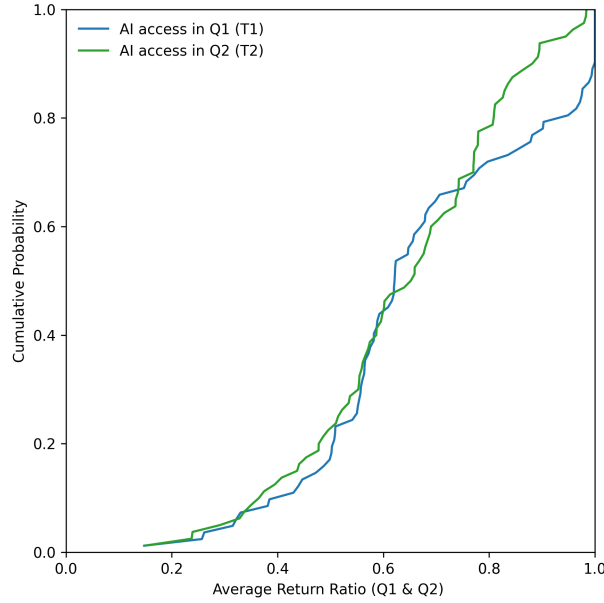


Figure 4: Average Return Ratio (Q1 & Q2)

who achieve near-perfect performance in both Q1 and Q2 (defined as an average performance above 0.99; 11 observations) are exclusively from the groups that received early AI access. This indicates that the performance advantage associated with early AI exposure carries over to subsequent tasks, at least for a subset of individuals.

Quantile regressions further confirm that this persistent effect is concentrated in the upper tail of the performance distribution. Results from quantile regression show that AI-first treatment has no statistically significant effect at the 10th, 25th, 50th, or 75th percentiles of Q1+Q2 performance. However, at the 90th percentile, the treatment effect is positive and statistically significant (coefficient = 0.118, $p = 0.000$). This pattern suggests that early AI exposure primarily increases the mass of top performers rather than shifting the entire distribution. In particular, high-performing subjects in Q1 appear able to internalize insights from AI interaction and sustain high performance even when AI is no longer available. ■

Finally, we compare performance when participants interact with AI for the first versus second time. Specifically, we compare participants receiving AI assistance in either Q1 or Q2 to participants receiving AI assistance again in Q3. We find substantial improvement on second exposure, suggesting that participants learn to interact with AI more effectively over time.

Result 7. *Performance increases significantly when participants interact with the AI for the second time rather than the first.*

Supporting Evidence. The share invested in the best fund is 65.04% among participants with AI assistance in Q3, compared to an average of 46.44% among participants receiving AI assistance for

the first time in Q1 or Q2. The corresponding shares invested in the top two funds are 77.92% and 66.78%. The differences are statistically significant (p-values 0.0006 and 0.0199). ■

Taken together, these results suggest that participants learn little from unaided exposure to the task itself, but can learn substantially from interacting with AI. The benefits of AI assistance therefore extend beyond immediate recommendations and include persistent improvements in subsequent decision-making.

3.3 Modes of Interaction

The dataset includes the full transcript of each participant’s interaction with the AI. Prior to running the experiment, we conjectured that most participant messages would fall into one of three primary categories: guidance-seeking, information-seeking, or recommendation-seeking.

Guidance-seeking messages include requests for advice on how to approach the task or explanations of relevant concepts. Information-seeking messages instead focus on extracting specific facts, such as asking for the costs of available funds or identifying the fund with the lowest fees. Finally, a message seeks a recommendation if asking, for instance, how to allocate funds (full delegation) or for a shortlist of funds to focus on. We preregistered these three categories together with three residual categories: other task-related interaction (e.g., using the AI to perform calculations relevant to the problem), unrelated interaction, and no AI interaction.¹⁴

We leveraged the same LLM (OpenAI’s o4-mini model) to categorize messages using this taxonomy. Each conversation transcript (of each participant) was submitted alongside a description of the categories, and the model was asked to identify all categories represented; Online Appendix B provides the prompt used.¹⁵ For a simple validation exercise, we provide in Online Appendix C all messages uniquely assigned by the model to each category, allowing readers to directly assess the accuracy and interpretability of the classification.

We find that the three main categories described above encompass nearly all participants’ messages to the AI. Just 9% of messages are classified as containing ‘other relevant’ or ‘irrelevant’ content. In addition, every interaction transcript contains at least one message identified as containing one of the three main categories, suggesting that the taxonomy captures the core substance of participant-AI interactions.

To better understand how participants interact with the AI, we compute for each conversation the relative prevalence of the three main preregistered interaction categories: guidance-seeking,

¹⁴We also preregistered a subdivision of the recommendation-seeking category based on whether sufficient information was provided to allow the AI to make the optimal recommendation. This distinction is not directly relevant for the present analysis and will instead be used further below.

¹⁵Since messages may contain multiple components corresponding to different categories, the model was instructed to report all overlapping categories.

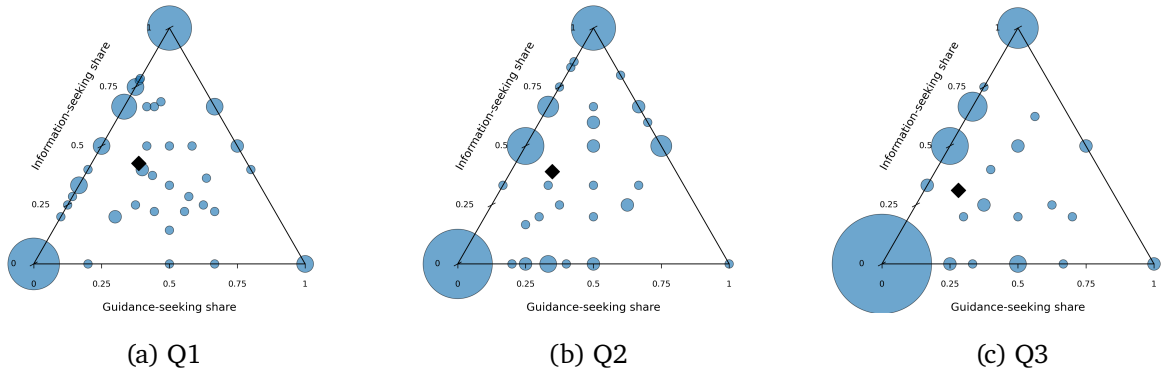


Figure 5: Participants’ interaction profiles in Q1, Q2 and Q3 (frequency weighted).

information-seeking, and recommendation-seeking. Each message is represented by a Boolean category vector (with a value of one in a category whenever the message intersects with it). The category vector for a conversation is the average vector over all messages the participant wrote during the conversation.

Figure 5 displays participants’ interaction profiles over the three main preregistered categories (that is, normalizing the conversation profile by the sum over these three categories). Each panel corresponds to one allocation task (Q1, Q2 or Q3). Points in the simplex represent distinct interaction profiles, with the size of the point reflecting the number of participants associated with that profile. The extreme vertices correspond to conversations concentrated on only one of these modes of interaction. We observe conversations clustered near each of these vertices, indicating that some participants rely almost exclusively on one interaction mode. At the same time, many observations lie in the interior of the simplex, reflecting conversations that combine multiple interaction modes. Overall, the figure reveals substantial heterogeneity in how participants interact with the AI, both along the extensive margin (which interaction modes are used) and the intensive margin (their relative prevalence within a conversation).

Result 8. *The vast majority (91%) of participants’ messages to AI fall in three main pre-registered categories: guidance-, information- and recommendation-seeking. Participant conversation profiles reveal substantial heterogeneity over these modes of interaction.*

3.4 Human Bottlenecks

As Result 2 shows, combining a simple prompt (full-information delegation) and a straightforward decision rule (adopting the LLM’s recommendation) yields essentially optimal performance. This benchmark highlights that in principle, AI can support highly efficient financial decisions. We find, however, that AI-assisted choices remain substantially below this frontier in practice (Result 3).

	Recommendation	Guidance	Information
Present	0.742 (0.029)	0.762 (0.033)	0.726 (0.029)
Absent	0.631 (0.046)	0.678 (0.037)	0.686 (0.048)
<i>p-value</i>	0.0517	0.0921	0.4706

Table 2: In the first column, average performance by presence or absence of recommendation-seeking messages in a subject’s first interaction with AI; and similarly for guidance-seeking and information-seeking in columns two and three. Standard errors in parentheses, with p-value from two-sided test of means in the bottom row.

The key question is therefore not whether AI can improve decisions, but which frictions prevent individuals from utilizing its full potential. We lay out and assess different potential explanations.

Possible bottleneck #1: mode of engagement. Recommendation-seeking is the least cognitively demanding mode of engagement (requiring only a simple prompt to obtain an optimal recommendation), but participants do not uniformly adopt it. As Result 8 shows, participants engage with the AI using a variety of interaction modes, including guidance-seeking and information-seeking. Such behavior is sensible if participants enjoy learning about the task, view the AI as a helpful tool rather than a decision-maker, or are skeptical of the AI’s recommendations. This last interpretation is consistent with our finding, discussed in Section 4, that participants substantially underestimate the precision of AI in identifying the optimal fund.

At the same time, as Table 2 shows, whether the subject’s interaction profile in their first AI-assisted task includes recommendation-seeking, guidance-seeking, or information-seeking, average performance remains substantially below that achievable via AI (Result 2), and suggests that simply engaging in recommendation-seeking is insufficient to eliminate the gap between observed and benchmark performance. We therefore turn to two additional sources of inefficiency, examining why average performance remains suboptimal regardless of interaction mode.

Possible bottleneck #2: suboptimal information provision to AI. Across interaction modes, AI’s ability to assist may be hampered if participants fail to properly describe their task. In particular, the quality of AI’s advice can hinge on whether it knows the key inputs (the investment horizon, the set of available funds, and the investment amount). The time horizon and available funds are especially important.¹⁶ Failing to specify the horizon when requesting a portfolio recommendation

¹⁶Failing to provide the investment amount is likely to be the least costly error for subjects, as the interface validates that the \$12,000 is spent and that the allocation meets minimum opening requirements. Subjects receive an error message if their selected allocation is not validated, and can simply enter another allocation.

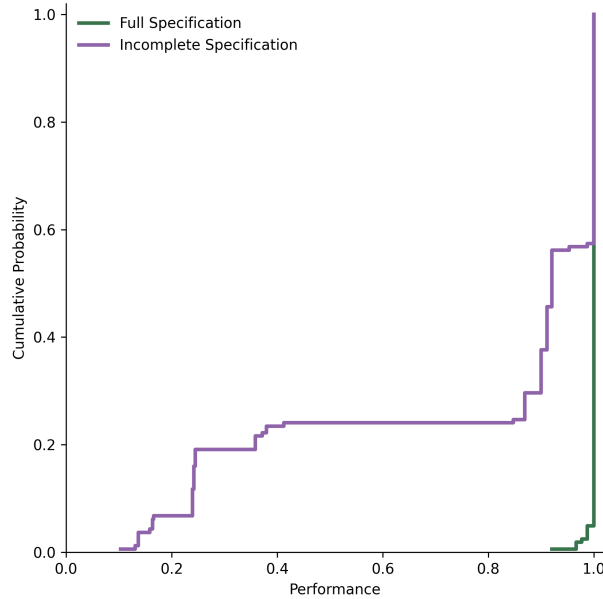


Figure 6: Simulated AI Performance by Completeness of Task Specification

may lead the AI to assume the horizon is longer and recommend funds with high one-time costs but low long-term costs. Similarly, suboptimal decisions and recommendations may occur if the menu is not clearly stated. Indeed, the user must indicate which tickers are available, as typical prospectuses (such as the ones we use) include not just the available fund, but a variety of share classes from the same firm that have different cost structures.

We evaluate the potential impact of suboptimal information provision through simulation. For each condition, we run 1,100 trials in which AI is asked to provide the return maximizing allocation. The *fully-specified delegation* benchmark is just as in Result 2: the prompt provides all relevant inputs to define the investment task (investment horizon, available funds, and investment amount) and the summary prospectuses. In the *incompletely-specified delegation* setting, the prompt is identical except that the investment horizon is omitted.¹⁷ The two CDFs of simulated performance in Figure 6 are generated under the above prompt structures, and differ only in whether investment horizon is included. The figure shows a stark performance gap between these two conditions. When the investment horizon is not provided, AI performance deteriorates substantially relative to the near-perfect benchmark achieved under full specification. This highlights that even a highly capable model cannot compensate for missing critical inputs.

We find that suboptimal information provision indeed contributes to the performance shortfall in our data. Among participants who engage in recommendation-seeking behavior during their

¹⁷Our simulation omits the horizon but specifies the index-fund tickers and total investment amount to ensure a feasible recommendation results (otherwise, assumptions would be needed regarding how participants respond to infeasible recommendations).

first AI-assisted task (i.e., either Q1 or Q2), only one-third fully specify the problem during their conversation with AI. This statistic improves, likely due to learning, during the second AI-assisted task (i.e., Q3). Yet even then, only 55%, or slightly more than half, provide all key inputs when seeking a recommendation. Improper specification leads to a large deterioration in performance for recommendation-seekers in Q1-Q3: a 25-point reduction in performance from about 0.927 to 0.676 (significant at all standard levels).

Result 9. *Suboptimal information provision to AI is a significant bottleneck. A substantial fraction of recommendation-seekers fail to specify key details of the portfolio choice problem, leading to a 25-point deterioration in their performance.*

Possible bottleneck #3: suboptimal information extraction. Even when interactions generate useful information, participants may not fully extract or apply it when making decisions. Participants who receive AI access can potentially use three sources of information: their conversation transcript with the AI, the task description, and the summary prospectuses for the four available funds. Decision quality depends not only on the total information contained in these sources, but also on the decision maker’s ability to extract and use that information effectively. This can be true even for the conversation with AI. For instance, participants may allocate only some of their budget to the suggested option, or struggle to process and synthesize the information when AI offers carefully detailed explanations. As a result, the quality of the decision may fall short of what would have been feasible given the information generated by the actual interaction.

We propose to quantify this channel through *AI-benchmarked information utilization*. This exercise disentangles suboptimal information provision to AI from suboptimal information extraction by the participant. The idea is to first assess an information source through simulation, by incrementally providing it to a capable AI and measuring the resulting performance gains. These simulated gains represent the best estimate for the full informational value contained in each source. Then, to evaluate how effectively *human* participants use these sources, we assess participants’ performance against the AI-extracted yardsticks.

Specifically, we prompt o4-mini under different informational environments and evaluate its success in simulations. The **Baseline environment (B)** provides only the four available fund tickers and a basic prompt asking for the return-maximizing allocation of \$12,000. The **Transcript (T)** environment additionally includes the participant’s full conversation with the AI. The **Horizon (H)** environment augments the Transcript environment by further including the investment horizon. The **Prospectus (P)** environment instead augments the transcript with the funds’ summary prospectuses.¹⁸ For each participant-question combination, and each informational condition, we

¹⁸We do not include both the prospectuses and the horizon, as this would be akin to the fully-specified delegation benchmark, under which we already know the AI achieves essentially perfect performance (even without the transcript).

compute the AI’s average performance across repeated simulations (using the same bundle of four funds and, when applicable, that participant’s transcript with AI). Performance in the Transcript, Horizon and Prospectus environments varies widely with the content of the conversation with AI; see the distributions in Figure B.11 in the Appendix. Notably, the distribution of performance in all these environments falls far below the near-perfect performance AI achieves in the fully-specified delegation benchmark.¹⁹

We use the performance gains between the simulated environments to quantify the *informativeness* of available sources:²⁰

- *Transcript Informativeness*: the difference between the AI’s average performance in the Transcript environment and the Baseline environment. This measures informational value of the conversation when read by a capable AI (as opposed to a potentially cognitively-constrained economic agent). An informative conversation should have strictly positive value. By contrast, a conversation that misguides the AI could have negative value. As can be seen from the computed distribution in Figure 7, the vast majority of transcripts (about 71%) have strictly positive informational value.
- *Incremental Horizon Informativeness*: the difference between the AI’s average performance in the Horizon environment and the Transcript environment. As both informational environments contain the conversation with AI, this captures the added value of taking into account the investment horizon, holding constant information already in the transcript. While the extent of performance gain may depend on the broader informativeness of the transcript, this measure is expected to be zero when the participant has already mentioned the horizon themselves. As visualized in Figure 7, adding the horizon has zero informational value in only about one-quarter of subject-question pairs.
- *Incremental Prospectus Informativeness*: the difference between the AI’s average performance in the Prospectus environment and the Transcript environment. Both informational environments include the conversation transcript, but only in the Prospectus environment simulation are the prospectuses uploaded to the AI. Importantly, the transcript from the original conversation does reflect an environment in which the summary prospectuses were uploaded. Hence the transcript may already include information drawn from the prospectuses, depending on

¹⁹In particular, the suboptimal performance in the Prospectus environment (which includes all information needed other than the horizon) offers yet another indication that the investment horizon is often not successfully communicated during the interaction, as discussed under possible bottleneck #2.

²⁰We assess measurement reliability of the simulated performance differences using a reliability ratio defined as $1 - \frac{\text{noise variance}}{\text{total variance}}$. The estimated reliability ratios are 0.94 for Transcript Informativeness, 0.88 for Incremental Horizon Informativeness, and 0.93 for Incremental Prospectus Informativeness, indicating that most variation reflects signal rather than simulation noise.

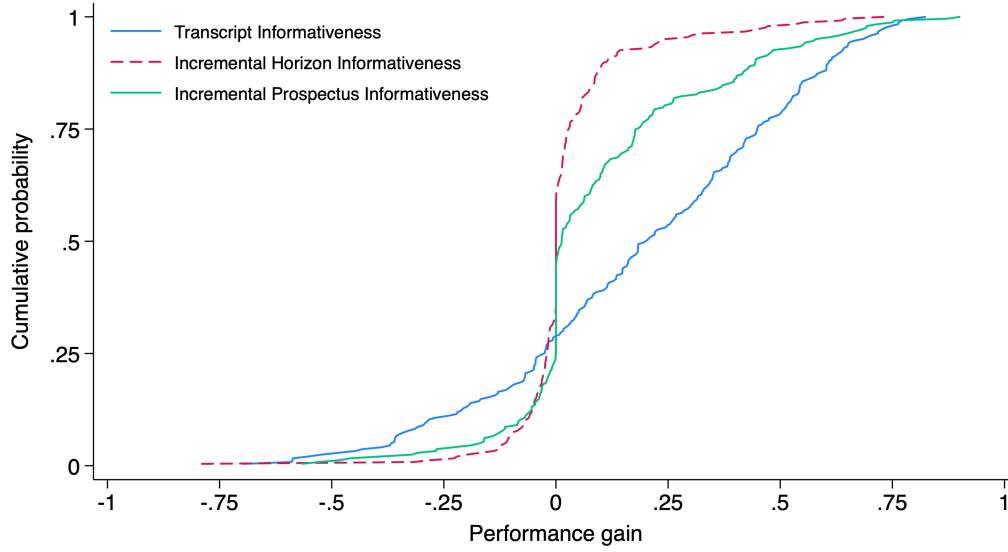


Figure 7: CDFs of informativeness measures computed at the subject-question level.

the prompts used by the subject. This incremental informativeness measure thus corresponds to the added value of capably synthesizing the information in prospectuses on top of information already appearing in the transcript. This measure is expected to be zero if the subject’s prompts led the AI to produce an optimal recommendation or to summarize the relevant information in the prospectuses during the conversation. Again from Figure 7, the incremental value of the prospectus is zero for about 25% of transcript-question pairs.

Table 3 presents the results of regressing participant performance on these AI-extractable measures of informativeness. Participant performance is positively and significantly associated with transcript informativeness, as well as with the incremental informativeness of the investment horizon on top of the transcript. The magnitudes of the AI-benchmarked informational utilization coefficients are also informative. The estimated coefficients on transcript informativeness and on the additional contribution of the investment horizon are 0.400 and 0.259, respectively, both well below one. This indicates that although participants benefit from these information resources, their performance gains are substantially smaller than those achieved by a capable AI given the same information. In other words, participants only partially extract and convert it into improved performance relative to a capable AI benchmark. By contrast, we find no evidence that participants are able to extract any additional useful information from the summary prospectuses beyond what is already reflected in the transcript, with the corresponding coefficient very small and not significantly different from zero. This finding confirms the intuition that summary prospectuses are difficult for participants to parse on their own, as they are lengthy, dense, and contain technical terminology.

Table 3: Estimates of AI-benchmarked information utilization

	Return Ratio
Transcript Informativeness $\times$ AI Access	0.400*** (0.052)
Incremental Horizon Informativeness $\times$ AI Access	0.259** (0.129)
Incremental Prospectus Informativeness $\times$ AI Access	-0.006 (0.088)
Constant	0.642*** (0.015)
Observations	486
Clusters (participant)	162
R^2	0.120

Notes: The dependent variable is the participant’s return ratio. All informativeness variables are interacted with an indicator for AI access. Standard errors are clustered at the participant level and reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Result 10. *Extracting information from the available resources is a significant bottleneck. Participants even have difficulty synthesizing their own conversation with AI: on balance, they are only able to extract about 40% of its AI-benchmarked informational value.*

4 Additional analyses

The richness of the dataset allows us to investigate several complementary questions relating to engagement with AI, trust in its capability, heterogeneity, and the mechanisms underlying the main results.

Question 1: How does AI impact decision-making effort and attention?

We find that the availability of AI assistance substantially shifts participants’ time and effort allocation. The first time participants face the portfolio problem provides a clean slate for establishing this. Recall from Section 3.1 that participants exert over nine minutes on Q1 when AI is unavailable. As has been shown theoretically by [Goncalves \(2026\)](#), the relationship between response times and task complexity may be non-monotone: when a difficult task is simplified, decision makers may find it worthwhile to spend more time on it. Our data suggests an effect along these lines when AI access is offered. On average in Q1, participants with AI access spend 5.55 more minutes on the task ($p = 0.000$). In the above comparison, the underlying portfolio choice problem is the same, but we interpret the availability of a helpful tool (AI) as potentially mitigating complexity.

At the same time, the nature of the effort exerted changes, as participants spend substantially less time engaging with the summary prospectuses. In principle, participants could have sought guidance from AI on how to read the prospectuses, with the goal of delving further into them. Instead, they spend less time on a prospectus per click ($p = 0.001$)²¹, and the mean file-viewing time decreases by approximately 45% (from 425.6 seconds without AI to 234.4 seconds with AI; $p = 0.000$). Further supporting this pattern, the proportion of subjects who open all four fund prospectuses within a question drops significantly from 88.1% to 45.3% when AI access is available ($p = 0.000$).

These patterns suggest that when AI assistance is available, participants engage more with the problem in a broad sense (via AI) but rely less on directly reading the prospectuses to acquire information. At least on that dimension, AI is used primarily to substitute for human effort rather than complement it. This supports the mechanism noted earlier, as the quality of the conversation with AI plays a critical role in obtaining the information to solve the task.

Question 2: Does access to AI increase confidence in own decisions, and do subjects perceive the AI to be accurate?

We assess every subject’s cognitive uncertainty (CU) immediately after Q1, asking them to self-report how sure they are that their chosen portfolio allocation is close to the optimal one. We also assess perceived accuracy in the Belief task using a binarized scoring rule, asking subjects to estimate the empirical rate at which AI correctly chooses the profit-maximizing fund in binary choice simulations.

Taken together, these measures reveal that while access to AI boosts confidence, subjects underestimate AI’s accuracy. Subjects have significantly more confidence in the optimality of their Q1 allocation when they have AI access, with the data displaying a first-order increase in the distribution of confidence. In particular, participants reported significantly higher average confidence when AI is available (46.71%) than when it is not (31.62%; $p = 0.000$). This increased confidence is justified in view of their improved performance with AI.

At the same time, subjects substantially underestimate AI’s ability to identify the optimal fund. The mean (and median) guess of AI’s success rate is 70%, well below its actual 95.45% success rate. Only 5% of subjects come close to the true rate with an estimate of 90% or more. More reassuringly, subjects generally believe that AI is more likely to be correct than not, with 85% estimating the success rate is strictly above 50%.

One might expect participants who perceive AI as more capable to be more confident that their AI-assisted portfolio is optimal. Consistent with this hypothesis, among participants with access to

²¹Time per click is calculated as total file-viewing time divided by the number of clicks used to open the files.

AI in Q1, confidence in the optimality of the chosen portfolio is positively correlated with subjects' estimates of AI's accuracy (Spearman correlation coefficient = 0.2634; $p = 0.0171$). The association is even slightly stronger among participants who sent at least one delegation message during their interaction with AI (Spearman correlation coefficient = 0.34; $p = 0.0066$).

Question 3: How does WTP for an optimal intervention depend on AI access?

Immediately following the portfolio allocation problem in Q3, we offer subjects in the WTP task the opportunity for an optimal intervention: their payoff in the WTP task could either be the profit from their chosen allocation in Q3 or, for a fee, the profit from the optimal portfolio in that problem. We elicit subjects' willingness to pay for this intervention using a multiple price list (MPL), with fees ranging from \$1 to \$25 in steps of \$2.²²

Recall that subjects in the 'plus' treatments had access to AI when choosing their portfolio in Q3, while subjects in the 'minus' treatments did not. Consistent with our results thus far, those with AI access in Q3 still achieve higher performance in that task. Indeed, in Q3, subjects with AI access achieve on average 80% of the optimal return, compared to 68% for subjects without it ($p = 0.0057$). None of the subjects had access to AI when answering the ensuing WTP question. We find that the average WTP of subjects with AI access in Q3 (\$10.37) is about 12.4% lower, but not quite significantly different from, the average WTP of subjects without it (\$11.84; $p = 0.1646$). This suggests that subjects do not fully recognize the improvement in decision quality gained from using AI. Such behavior may arise, for instance, if subjects follow AI's recommendation but are unable to assess the recommendation's quality.

Both subjects with and without AI assistance in Q3 overvalue the optimal intervention in the WTP task. The average overpayment is large: \$4.25 for subjects with AI access in Q3, compared to \$2.49 by subjects without it; both overpayments are significantly different from zero, with the largest p-value 0.0091. The median WTP is substantial even among subjects who selected an optimal portfolio in Q3 (\$10 with AI, \$8 without). A possible explanation is that subjects assess their Q3 performance with caution, reflecting a form of ambiguity aversion.

A fuller picture emerges when investigating how WTP varies with actual Q3 performance. We find that the relationship between performance and WTP for the intervention is significantly weaker for subjects who had AI access in Q3. Figure 8 illustrates this by plotting the predicted WTP as a

²²MPLs can be challenging for subjects, with data sometimes showing multiple switching points, or more generally being willing to pay a higher fee but not some smaller one. Our interface includes text describing each option to minimize confusion, and subjects must fill in each row independently. Reassuringly, nearly all our subjects behave in a consistent manner: only 6 out of 162 subjects make choices that prevent us from eliciting a well-defined WTP. These 6 subjects are thus excluded from the WTP analysis. Following standard practice, we define WTP as the midpoint of the grid around the switch point (e.g., a subject willing to pay \$23 but not \$25 is assigned a WTP of \$24). For a subject who isn't willing to pay even the lowest listed fee of \$1, we assign a WTP of \$0.50. For a subject willing to pay all listed fees, we top-code a WTP of \$25.

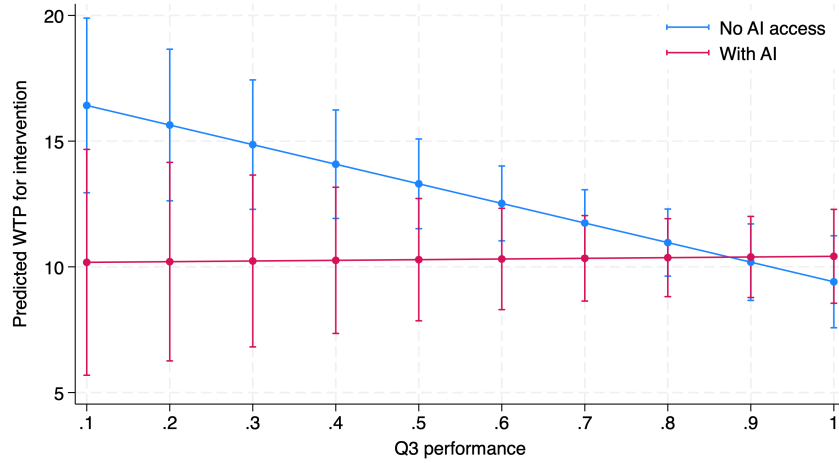


Figure 8: Predicted willingness to pay for the optimal intervention for subjects with and without AI access in Q3, based on linear regression of WTP on performance, AI access, and their interaction. Bands represent 95% robust confidence intervals.

function of Q3 performance separately for each group of participants. For those without AI access, a 10-point increase in Q3 performance is associated with about a 78 cent reduction in WTP for intervention ($p = 0.002$). With AI, we find no statistically significant relationship between WTP and performance ($p = 0.929$).²³

Thus, while subjects correctly expect AI assistance will improve their performance (as shown in Question 2), they can better gauge the quality of their decisions when constructing allocations independently. One might instead have expected better-calibrated WTP from subjects with AI access in Q3, as they had an additional opportunity to learn about the task from AI. This suggests that there is a distinction between having increased confidence that one has made an optimal choice, versus increased confidence in one's ability to construct the monetary valuation of that choice.

Question 4: Who benefits most from AI access?

Our optional exit survey collects self-assessments of financial knowledge and experience with AI prior to the survey, with the response to each question recorded using a 5-point Likert scale.²⁴ Though subjective and coarse, these self-assessments may help us better understand how AI may mediate the benefits of prior financial knowledge and AI experience. In particular, we conjectured that participants with lower financial literacy and higher prior AI experience will benefit most from

²³The same pattern arises when controlling for estimated accuracy of AI, Q3 response time and prospectus engagement in Q3; see the regression output in Table B.7 in the Appendix.

²⁴The survey asks, "Before starting this experiment, how knowledgeable did you consider yourself to be about financial investments?" and "Before starting this experiment, how experienced an AI user did you consider yourself to be?" Qualitative descriptions of responses range from 'not at all' to 'extremely' knowledgeable or experienced.

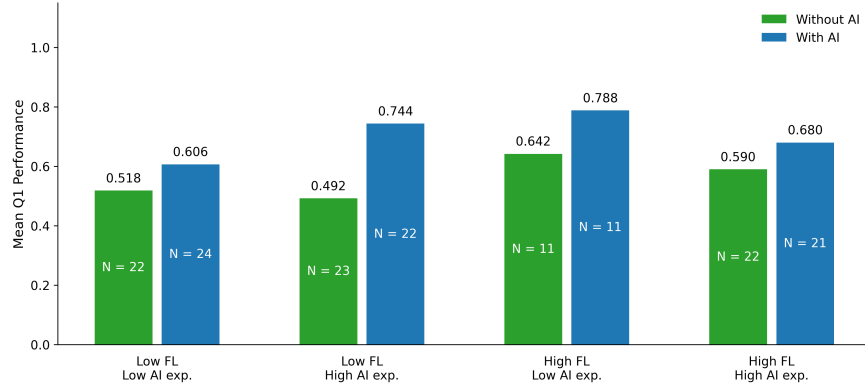


Figure 9: Performance Gains by Financial Literacy and Prior AI Experience

AI. To study this, we focus on the first allocation task to abstract from learning due to repeated task exposure. On each dimension (prior AI use and financial literacy) we divide the sample as evenly as possible into ‘high’ and ‘low’ groups. Figure 9 displays average performance for each combination. We estimate a model allowing the impact of AI assistance to differ between participants exhibiting the combination of high AI experience and low financial literacy and other participants. Though there is some suggestive evidence for our conjectured hypothesis, in the sense that the mean increase appears largest for this group, the difference in the AI effect between this group and other participants does not quite reach statistical significance ($p = 0.137$). Broadly, what we do clearly see from the Figure is that even when AI is available, participants in all categories perform well below AI’s capabilities, suggesting they make inefficient use of the tool.

Motivated by this, we consider whether the self-reported scores for prior AI experience and financial literacy are associated with particular modes of AI engagement. We find that greater self-reported financial literacy predicts significantly more information-seeking at the expense of less guidance-seeking. In view of their low performance even with AI, it thus appears that more financially-literate participants overestimate their understanding of the investment problem. By contrast, participants with greater self-reported prior exposure to AI are marginally more likely to seek recommendations rather than retrieve information. Table B.8 in the Appendix provides details. These findings provide some suggestive evidence that interaction modes are not entirely fixed, but adapt to participants’ backgrounds. Having said this, greater variation in LLM experience and financial literacy are needed to draw more robust conclusions, suggesting an interesting area for future research.

5 Conclusion

Unlike technologies tailored to a specific environment, large-language models offer an all-purpose tool that can potentially alleviate the cognitive constraints which impact decision quality in a wide range of tasks. Given their generality, these tools rely on open-ended, organic interaction with the user rather than pre-determined inputs. Our findings suggest that this generality comes at a potential cost: even when the LLM is highly capable in a task, the benefit provided by the technology depends greatly on how humans interact with it. In our setting, human performance is substantially improved by access to AI, yet remains well below the level the AI itself is capable of achieving (when given the information needed to solve the problem). This shortfall arises, in part, because humans may fail to convey important details of the task and have difficulty utilizing the information in AI's response. From this perspective, AI-augmented choice is a joint-production process in which some cognitive constraints are mitigated but additional constraints, namely those at the human-AI interface, are introduced.

One possible response to these findings is to ask why the human should remain involved at all: if AI can make nearly optimal investment decisions, why not simply delegate the decision to AI? Our experimental design and results highlight both a practical and a conceptual limitation of such an approach. However capable an LLM may be, it can only solve the task that is conveyed. In our setting, critical information includes the investment horizon and available index funds – details which participants often fail to communicate. An automated decision-making system requires access to these details, which can include varying constraints and objectives that are not immediately observable. Furthermore, individuals may value some degree of autonomy, not least because their personal preferences and context may impact the optimal decision and must be conveyed. We abstract from personal preferences in our experimental design, as the solution to the portfolio choice problem is deliberately common across all subjects to obtain a well-defined benchmark for decision quality. It would be important for future work to also explore the questions we study in a setting with subjective preferences.

Our findings do suggest that interface design could play an important role in addressing human bottlenecks. Human information utilization of AI's responses could potentially improve if the AI provides short and simple responses that minimize attentional burden; at the same time, this may obscure important nuance. Decision quality in our setting may have also improved had the AI tool proactively prompted users for essential details of the environment or communicated evidence of its capabilities to promote recommendation-seeking. Again, there are nontrivial costs and benefits: on the one hand, an AI that does not prompt the user to fill in details must make potentially incorrect assumptions to fill the gaps; on the other hand, a high interaction burden may discourage users by reducing the cognitive savings from using the tool. How to navigate these tradeoffs offers a

fascinating direction for research on AI as a tool for economic decision making.

References

- AGARWAL, N., A. MOEHRING, P. RAJPURKAR, AND T. SALZ (2023): “Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology,” Working Paper 31422, National Bureau of Economic Research.
- AGARWAL, N., A. MOEHRING, AND A. WOLITZKY (2026): “Designing Human-AI Collaboration: A Sufficient-Statistic Approach,” Working paper, August 20, 2026.
- ANGELOVA, V., W. DOBBIE, AND C. S. YANG (2026): “Algorithmic Recommendations and Human Discretion,” *The Review of Economic Studies*, 93, 2250–2283.
- BESHEARS, J., J. J. CHOI, D. LAIBSON, AND B. C. MADRIAN (2011): “How Does Simplified Disclosure Affect Individuals’ Mutual Fund Choices?” in *Explorations in the Economics of Aging*, ed. by D. A. Wise, University of Chicago Press, 75–96.
- BRYNJOLFSSON, E., D. LI, AND L. RAYMOND (2025): “Generative AI at Work,” *The Quarterly Journal of Economics*, 140, 889–942.
- CAPLIN, A. AND M. DEAN (2015): “Revealed Preference, Rational Inattention, and Costly Information Acquisition,” *American Economic Review*, 105, 2183–2203.
- CAPLIN, A., D. DEMING, S. LI, D. MARTIN, P. MARX, B. WEIDMANN, AND K. J. YE (2025): “The ABCs of Who Benefits from Working with AI: Ability, Beliefs, and Calibration,” *Management Science*, 72, 5843–5852.
- CHEN, M. H., E. FILIZ-OZBAY, AND E. Y. OZBAY (2026): “The Impact of LLM-Based Recommendations on Decisions under Uncertainty,” Working paper.
- CHOI, J. H. AND D. SCHWARCZ (2025): “AI Assistance in Legal Analysis: An Empirical Study,” *Journal of Legal Education*, 73, 384–420.
- CHOI, J. J., D. LAIBSON, AND B. C. MADRIAN (2010): “Why Does the Law of One Price Fail? An Experiment on Index Mutual Funds,” *The Review of Financial Studies*, 23, 1405–1432.
- CHOUKHMANE, T., T. DE SILVA, W. LIN, AND M. AKUZAWA (2026): “AI Financial Advice: Supply, Demand, and Life Cycle Implications,” Working Paper 35574, National Bureau of Economic Research.
- CONLON, J. J. AND P. SCHWARDMANN (2026): “AI Sycophancy and Decisions,” Working paper, April 17.
- DE CLIPPEL, G., P. MOSCARIELLO, P. ORTOLEVA, AND K. ROZEN (2025): “Caution in the Face of Complexlity,” *Working paper*.
- DE CLIPPEL, G. AND K. ROZEN (2024): “Bounded Rationality in Choice Theory: A Survey,” *Journal of Economic Literature*, 62, 995–1039.
- DIETVORST, B. J., J. P. SIMMONS, AND C. MASSEY (2015): “Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err,” *Journal of Experimental Psychology: General*, 144, 114–126.

- DREYFUSS, B. AND R. RAUX (2026): “Human Learning about AI,” Working paper, revised May 2026.
- ENKE, B. (2024): “The Cognitive Turn in Behavioral Economics,” Working paper.
- ENKE, B. AND T. GRAEBER (2023): “Cognitive Uncertainty,” *The Quarterly Journal of Economics*, 138, 2021–2067.
- FEDYK, A., A. KAKHBOD, P. LI, AND U. MALMENDIER (2026): “AI and perception biases in investments: An experimental study,” *Journal of Financial Economics*, 185, 104350.
- GABAIX, X. (2014): “A Sparsity-Based Model of Bounded Rationality,” *The Quarterly Journal of Economics*, 129, 1661–1710.
- GONCALVES, D. (2026): “Speed, Accuracy and Complexity,” Working paper.
- GRUBER, J., B. R. HANDEL, S. H. KINA, AND J. T. KOLSTAD (2021): “Managing Intelligence: Skilled Experts and AI in Markets for Complex Products,” *RAND Journal of Economics*, forthcoming.
- HOONG, R. AND B. DREYFUSS (2026): “Calibrated Coarsening: Designing Information for AI-Assisted Decisions,” Working paper, revised July 2026.
- IMAS, A., K. LEE, AND S. MISRA (2026): “Agentic Interactions,” Working paper.
- KLEINBERG, J., H. LAKKARAJU, J. LESKOVEC, J. LUDWIG, AND S. MULLAINATHAN (2018): “Human Decisions and Machine Predictions,” *The Quarterly Journal of Economics*, 133, 237–293.
- KLEINBERG, J., J. LUDWIG, S. MULLAINATHAN, AND Z. OBERMEYER (2015): “Prediction Policy Problems,” *American Economic Review*, 105, 491–95.
- MCLAUGHLIN, B. AND J. SPIESS (2024): “Designing Algorithmic Recommendations to Achieve Human-AI Complementarity,” Working paper.
- NOY, S. AND W. ZHANG (2023): “Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence,” *Science*, 381, 187–192.
- OPREA, R. (2026): “Complexity and Its Measurement,” in *Handbook of Experimental Methods in the Social Sciences*, ed. by A. Rees-Jones, Edward Elgar, chap. 3, 104–161.
- OTIS, N. G., R. CLARKE, S. DELECOURT, D. HOLTZ, AND R. KONING (2026): “The Uneven Impact of Generative Artificial Intelligence on Entrepreneurial Performance: Evidence from a Field Experiment in Kenya,” *Management Science*, published online.
- SALANT, Y. (2011): “Procedural Analysis of Choice Rules with Applications to Bounded Rationality,” *American Economic Review*, 101, 724–48.
- SIMON, H. A. (1976): “From Substantive to Procedural Rationality,” in *Method and Appraisal in Economics*, ed. by S. J. Latsis, Cambridge University Press, 129–148.
- SIMS, C. A. (2003): “Implications of Rational Inattention,” *Journal of Monetary Economics*, 50, 665–690.
- STIGLER, G. J. (1961): “The Economics of Information,” *Journal of Political Economy*, 69, 213–225.

Vafa, K., A. Rambachan, and S. Mullaianathan (2024): “Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function,” in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, vol. 235 of *Proceedings of Machine Learning Research*, 48919–48937.

APPENDIX

A Additional Details

The twelve S&P 500 index funds used in the experiment are presented in Table A.4.

Table A.4: Summary of the 12 S&P 500 Index Funds Used in the Experiment

Ticker	Fund Name	Load (%)	Expense Ratio (%)	Minimum Opening (\$)
OEICX	JPMorgan Equity Index Fund Class C	1.00	1.13	1,000
WFINX	Allspring Index Fund Class C	1.00	1.19	1,000
SXPCX	DWS S&P 500 Index Fund Class C	1.00	1.26	1,000
SPICX	Invesco S&P 500 Index Fund Class C	1.00	1.28	1,000
MUXAX	Victory S&P 500 Index Fund Class A	2.25	0.54	2,500
RYSYX	Rydex S&P 500 Fund Class C	1.00	2.36	2,500
SXPAX	DWS S&P 500 Index Fund Class A	4.50	0.55	1,000
OGEAX	JPMorgan Equity Index Fund Class A	5.25	0.45	1,000
SPIAX	Invesco S&P 500 Index Fund Class A	5.50	0.54	1,000
WFILX	Allspring Index Fund Class A	5.75	0.44	1,000
GRMAX	Nationwide S&P 500 Index Fund Class A	5.75	0.58	2,000
RYSOX	Rydex S&P 500 Fund Class A	4.75	1.60	2,500

Note: Subjects were provided with the most recent summary prospectuses available at the beginning of the investment period. The table summarizes the information contained in the summary prospectuses accessible to participants during the experiment.

Table A.5 and A.6 check for balance between treatment groups.

Table A.5: Mean Fund Returns by Treatment Group: Optimal Return

Group	Q1 Funds Optimal Return	Q2 Funds Optimal Return	Q3 Funds Return
T1-	1.049482	1.049922	1.049833
T1+	1.049823	1.050354	1.049673
T2-	1.049371	1.050019	1.049962
T2+	1.049744	1.0499	1.049608
Total	1.049606	1.05005	1.049769

Figure A.10 validates the cost minimization strategy by plotting the net return against total annual cost for each year from 2015 to 2024 for the twelve S&P 500 index funds used in our

Table A.6: Mean Fund Returns by Treatment Group: Average Return

Group	Q1 Funds Avg. Return	Q2 Funds Avg. Return	Q3 Funds Avg. Return
T1-	1.028112	1.028128	1.028315
T1+	1.027855	1.028448	1.028252
T2-	1.027768	1.028323	1.028464
T2+	1.027946	1.028649	1.02796
Total	1.027921	1.028386	1.028248

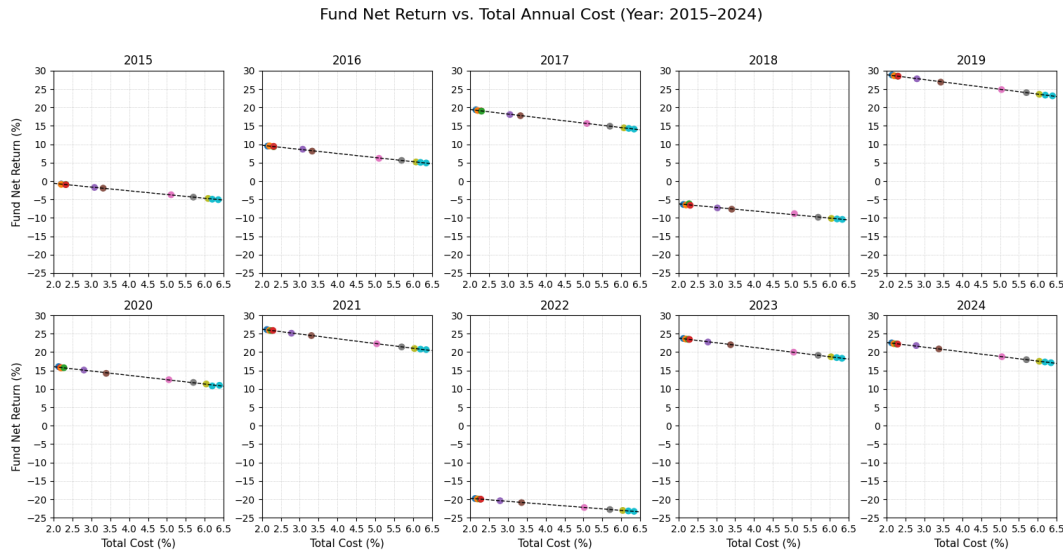


Figure A.10: Net Return Plotted against Total Annual Cost for the 12 Funds from Year 2015 to 2024.

Notes: Each panel plots the 12 funds used in a given year, with each colored dot representing a single fund. Each color tracks a fund across years. The vertical axis reports each fund’s net return (%) for that year—that is, the actual return received by investors after accounting for total annual costs. The horizontal axis reports each fund’s total annual cost (%).

experiment. In every year, we observe a consistent negative relationship between total cost and net return, confirming that lower-cost funds systematically deliver higher net returns.

B Additional Results

We include here additional figures and tables referenced in the main text.

Table B.7: Willingness to Pay (WTP) on Return Ratio and AI Access

	(1) WTP	(2) WTP
Return ratio	-7.792*** (2.533)	-7.780*** (2.548)
AI Access	-7.046** (3.240)	-2.278 (4.766)
AI Access $\times$ Return ratio	8.058** (3.896)	7.877** (3.924)
AI access $\times$ Belief of AI accuracy		-0.068 (0.053)
Total time		0.001 (0.003)
File view time		-0.000 (0.005)
Constant	17.199*** (1.995)	17.017*** (2.210)
Observations	156	156
R-squared	0.063	0.074
F-statistic	3.50**	2.08*

Note: Robust standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

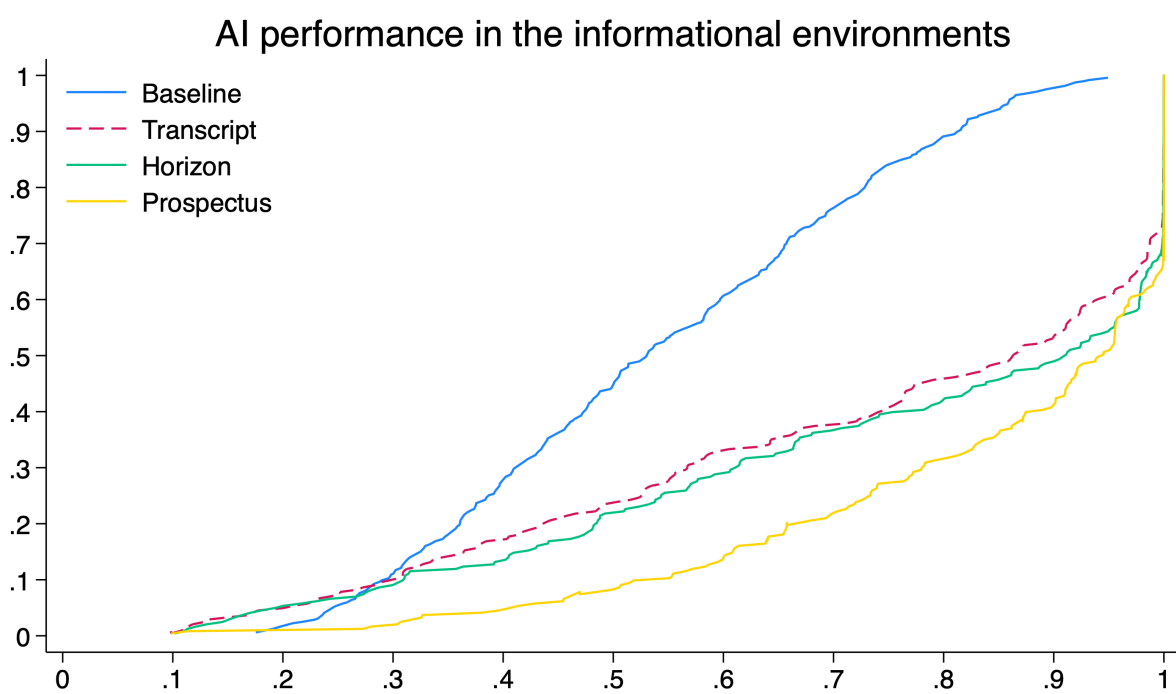


Figure B.11: AI Performance Benchmarks

Table B.8: Average Marginal Effects on Interaction Shares in Q1

	Recommendation-Seeking	Guidance seeking	Information seeking
AI experience	0.078 (0.050)	0.006 (0.036)	−0.083* (0.047)
Financial literacy	−0.031 (0.045)	−0.168*** (0.058)	0.199*** (0.034)
Observations	70	70	70

Notes: Average marginal effects from a fractional multinomial logit model estimated on Q1 observations. Standard errors clustered by participant are reported in parentheses. Effects are expressed as changes in interaction shares on a 0–1 scale. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.