Theorizing with Large Language Models

Matteo Tranchero*
University of Pennsylvania
mtranc@wharton.upenn.edu

Cecil-Francis Brenninkmeijer
University of California, Berkeley
cecil-francis.b@berkeley.edu

Arul Murugan
University of California, Berkeley
arul@berkeley.edu

Abhishek Nagaraj*†
University of California, Berkeley
nagaraj@berkeley.edu

July 31, 2025

Abstract

Large Language Models (LLMs) are emerging as powerful tools for management and organizational research. While early applications have centered on data processing and the replication of survey-based studies, their potential for theory building remains largely untapped. In this paper, we show how AI can support theory development by simulating experiments using LLMs as synthetic subjects. We begin by introducing a structured framework to standardize the design and execution of in silico experiments with LLM-powered agents. Although not without limitations, we argue that this approach blends the realism of human-subject data with the computational control of agent-based models. We apply the framework to the classic exploration–exploitation dilemma in strategy, showing that LLM simulations can reproduce "ground truth" results from human participants. We then demonstrate how LLM agents can be used to generate new hypotheses, investigate mechanisms, and test boundary conditions. We conclude by discussing the promise and limitations of this methodological innovation, and how it can complement existing approaches while opening new avenues for management research.

Keywords: AI Agents, Large Language Models, Strategy Theorizing, Strategic Interactions, Exploration and Exploitation

*Equal contribution

†Corresponding author. We thank participants at the Macro Research Lunch (UC Berkeley-Haas), 2024 Academy of Management, Faculty Research Lunch (UC Berkeley-Haas), FOOLs (Wharton), HIVA seminar (KU Leuven), 2024 Conference on AI, Machine Learning and Business Analytics (Yale SOM), AI & Strategy Seminar, GenAI Lab seminar (TUM School of Management), AI and Entrepreneurship Workshop (HEC Paris), 2025 Utah Winter Strategy Conference for useful feedback. We acknowledge support from OpenAI in the form of computing credits to use their GPT class of models. All errors are our own.

1 Introduction

Whether practiced by managers or studied by scholars, strategy hinges on understanding how actions translate into competitive outcomes in dynamic environments (Leiblein et al., 2018). Just as researchers investigate the returns from research and development (R&D), company executives want to know whether their innovation strategy will boost sales and how competitors are likely to respond. But this is easier said than done. In most business settings, multiple actors shape outcomes, and their decisions are often interdependent (Rumelt et al., 1991; Siggelkow, 2011; Van den Steen, 2018). Consider a pharmaceutical firm deciding which drug to develop. Sticking with known molecules may yield predictable but modest returns, while exploring new compounds involves substantial risks. And even if the risky bet pays off, rivals may quickly follow with cheaper alternatives that draw customers away—ultimately shrinking, not growing, profits.

Strategic interdependencies make it difficult to trace clear causal links between actions and outcomes. The ideal way to isolate mechanisms in such complex settings is through real-world and long-run experiments at the industry level, but these are rarely feasible (Adner and Levinthal, 2024). As a result, researchers usually turn to two alternative approaches to run strategy experiments. The first relies on computational models to simulate scenarios including nonlinear effects and feedback loops, interactions that closed-form models struggle to capture (Davis et al., 2007). These models offer tight control, enabling researchers to tweak key parameters and observe their impact on emergent dynamics (Knudsen et al., 2019). The second approach involves human subjects experiments, either in the laboratory or in partner organizations (Di Stefano and Gutierrez, 2019; Levine et al., 2023). This method ensures realism by capturing how actors behave under specific conditions, while holding constant other strategic interactions to isolate causal effects (Chatterji et al., 2016).

Experiments using either agent-based simulations or human subjects both offer distinct advantages, but both approaches also entail trade-offs between control and realism. On the one hand, the clean abstraction provided by simulated agent-based models (ABMs) helps isolate key dynamics while increasing the distance from real-world scenarios (Ganco, 2017). The very strength of ABMs is what complicates translating their insights to managerially-relevant situations (Knudsen, 2024). For example, ABMs excel in identifying feedback loops and interactions, but cannot provide estimates on the real-world impact of any strategic decision. On the other hand, lab and field experiments generate data from real-world behavior. But they come at a much higher cost and limited flexibility in examining multiple alternate scenarios (Bandiera et al., 2011; Davis et al., 2007).

Recent advances in generative artificial intelligence (AI) herald a new approach that could strike a different balance between control and realism in experiments. Large Language Models (LLMs), trained on vast corpora of human text, can emulate human language with remarkable fluency. This capability also enables them to approximate human behavior to a reasonable degree (Horton, 2023; Tu et al., 2024). As a result, scholars have begun to deploy LLMs as proxies for human behavior in fields such as computer science and computational social sciences—an approach referred to as "silicon sampling" (Dillion et al., 2023). Moreover, as software programs, LLMs can be instructed to assume a wide range of roles and be embedded in fully manipulable virtual environments. This opens the door to creating ecologies of realistic AI agents that interact within contexts mimicking the strategic interdependencies of real-world settings, offering a potentially valuable tool for both research and practice.

In this paper, we propose a general framework for using AI agents as synthetic subjects in strategic management. Our approach is inspired by the use of model organisms in the biomedical sciences: species sharing enough biological similarity to humans that allow researchers to refine initial theories before advancing to human testing. We then apply our framework to a simple exploration–exploitation problem involving strategic interdependencies. First, we show that our simulated experiments reproduce outcomes from experiments using human subjects, suggesting a promising degree of realism for AI agents. Second, we demonstrate how the substantial control LLMs provide over relevant variables enables us to rapidly explore extensions, facilitating the theorizing process.

In our framework, theorizing begins with human strategists defining a "starting theory" (Davis et al., 2007) or "thesis" (Ott and Hannah, 2024). This starting point identifies potential relationships between actions and outcomes worth exploring, while also defining the relevant payoffs and objectives (Agrawal et al., 2019). Once the theoretical space has been outlined, LLMs offer a powerful tool for flexible and rapid in-silico experimentation (Manning et al., 2024). Specifically, strategists can instantiate multiple LLM-powered agents with non-deterministic behaviors and embed them in complex strategic environments. Through simulated experiments, strategists can manipulate actions and parameters, observing agent interactions to generate predictions about the strength and direction of causal relationships. Just as laboratory mice occupy an intermediate position between in vitro testing and clinical trials (LaFollette and Shanks, 1995), AI agents can serve as a complementary tool to ABMs and human subject experiments.

LLM-powered simulations bring both opportunities and challenges. A key strength of LLMs lies in their ability to adopt diverse personas and demonstrate nuanced capabilities, often through simple priming with additional scripts. This gives AI agent simulations a level of on-the-ground plausibility that is difficult to achieve in traditional ABMs. Moreover, techniques like chain-of-thought elicitation and model interpretability allow researchers to peer into the "brain" of the decision-maker, however imperfectly. Even when LLM agents merely approximate human responses, the method retains all the core advantages of computer simulation: it is fast, low-cost, and easily repeatable, allowing researchers to explore a wide range of hypotheses and moderators at scale.

However, this approach also has important limitations along the very dimensions of realism and control that make it appealing. On the realism front, LLMs can exhibit a rationality bias, deviating from the bounded rationality typical of human decision-making (Hagendorff et al., 2023). LLMs are also prone to hallucination, meaning their stated motivations may not reflect genuine underlying mechanisms. As a result, LLM outputs must be interpreted with care by the researcher and cannot always be taken at face value as faithful representations of human behavior. In terms of control, LLMs remain imperfect: they are ultimately black boxes, making it difficult to precisely understand or manipulate their decision-making in the way traditional ABMs allow. In the end, we argue that this method occupies a novel position on the trade-off frontier between realism and control, offering a potentially powerful complement to existing approaches in strategy research.

With these pros and cons in mind, we apply our framework to the canonical problem in strategy: the exploration-exploitation dilemma (March, 1991). In many settings, firms must choose among risky options with uncertain value—such as a manager deciding which technologies to pursue. We revisit a recent study suggesting that when individual exploration generates public data for competitors, firms may overly focus on exploiting known options to their own detriment (Hoelzemann et al., 2025). This theory was initially tested in the laboratory with undergraduate participants. When we adopt our framework, simulating the same experiment using LLMs in place of human subjects, we observe comparable results. We then extend the original study by systematically varying the conditions of the original experiment. In doing so, we uncover new boundary conditions that clarify when the theory's predictions are likely to fail, as well as novel mechanisms that explain why they might persist. Notably, our findings were not anticipated by the LLM, suggesting that the outcome of the simulation was ex ante unknown even to the model itself (Manning et al., 2024). This application offers a first example of the validity and potential of our approach.

Our paper makes several contributions. First, we introduce a new tool for abductively building strategy theory (Mueller, 2018). We show that the combination of realism and control offered by experiments with AI agents is especially generative for theorizing about novel mechanisms. While the high costs of human experiments limit their potential, we show how AI agents can be leveraged as synthetic subjects to run openended simulations at scale. Second, we highlight both the potential and the limitations of AI simulators for strategy research, building on the rapidly growing literature on AI simulators in computer science (Aher et al., 2023; Dillion et al., 2023; Horton, 2023; Park et al., 2023). On one hand, the black-boxed engines of LLM-powered agents enable unexpected discoveries, similar to how machine learning is used to explore data (Choudhury et al., 2021; Shrestha et al., 2021). On the other hand, AI agents fall short of the realism of real-world experiments (Broska et al., 2024), and they lack the full control offered by simpler ABMs (Tjuatja et al., 2023). As such, we argue that LLMs should complement—rather than replace—the other tools in the strategist's toolkit. Finally, our application to the dynamics of exploration and exploitation demonstrates how LLM-powered experiments can lead to substantial theoretical contributions in practice. The new mechanisms and boundary conditions we uncovered offer intriguing hypotheses for future testing.

2 Strategy Theorizing With LLMs

2.1 Does Strategy Research Need a New Tool?

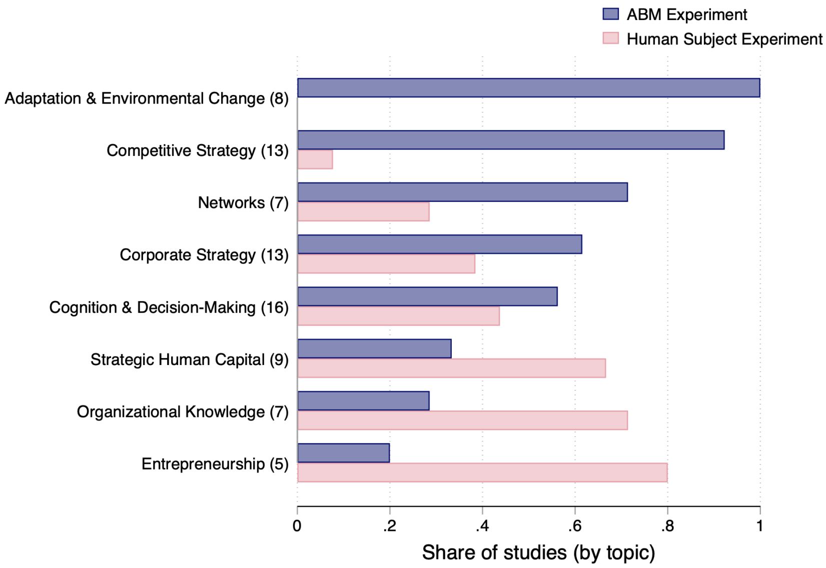
Strategic management aims to understand business dynamics and offer normative guidance to improve firm performance (Rumelt et al., 1991). A central challenge, however, is that strategic choices are often deeply interconnected, making it hard to pinpoint which mechanisms drive outcomes. While randomized controlled trials (RCTs) remain the gold standard, implementing them in contexts with actual business decisions is usually unfeasible (Adner and Levinthal, 2024). As a solution, strategy research is distinctive in its reliance on two experimental methods: in-silico experiments with agent-based models (ABMs)[1] and experiments with human subjects either in the lab or the field (Chatterji et al., 2016; Di Stefano and Gutierrez, 2019; Knudsen et al., 2019). To better understand how these methods are used in practice, we conducted a systematic review of studies appearing in the *Strategic Management Journal* (SMJ)—the field's longest-running and most representative outlet. Using Scopus, we identified 81 relevant articles published between 1980 and 2025: 49 using ABMs and 32 employing human subjects experiments (details in Appendix A). Common topics across these papers include cognition and decision-making, competitive and corporate strategy, strategic human capital, adaptation, and knowledge-related themes.[2]

[1] While mathematical models are used in strategy research (e.g., Van den Steen 2018), they often fall short in capturing the complexity and heterogeneity inherent in strategic interactions (Harrison et al., 2007; Knudsen, 2024). To overcome these limitations, scholars frequently turn to computational simulations when analytical solutions become intractable. However, the distinction between mathematical and computational models is largely one of degree: both rest on the same epistemic logic, but simulations offer greater flexibility to expand state spaces and capture richer patterns of interaction (Knudsen et al., 2019).

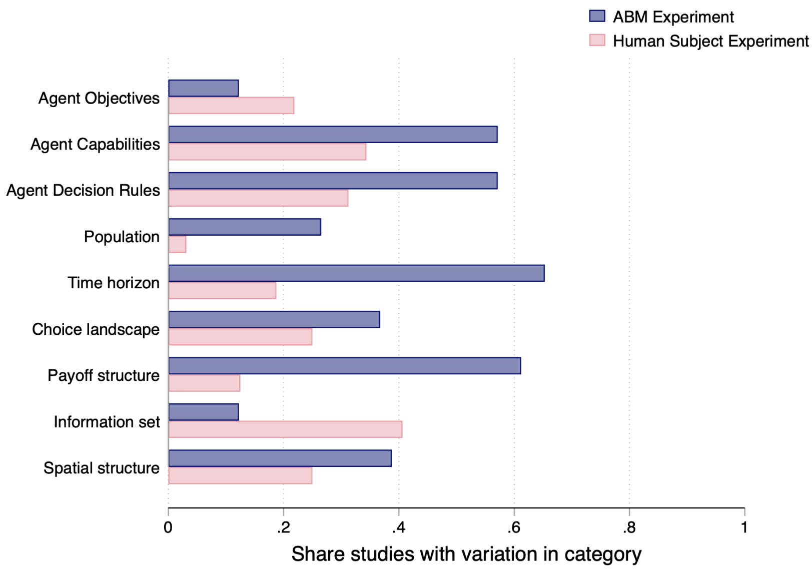
[2] Unlike the field of economics, strategy lacks a standardized classification system such as the *Journal of Economic Literature* (JEL) codes to sort articles by topic. For the purposes of this article, we inductively developed thematic categories to capture each study's primary focus. These categories are not intended to be universal or definitive, but rather to give a general sense of where ABMs and human experiments tend to be used in strategy research. The full database of the 81 studies we analyzed is available in the Supplementary Material, allowing readers to transparently replicate our analysis or use alternative categorizations.

Figure 1: Agent-Based and Human Subject Experiments Published in the *Strategic Management Journal* (1980-2025)

Panel A: Research Topics Studied by Methodological Approach



Panel B: Types of Experimental Manipulation by Methodological Approach



Note: This figure summarizes our review of articles published in the *Strategic Management Journal* that use either simulation-based or human subject experiments. Panel A shows the proportion of papers in each research topic by methodological approach. Panel B presents the proportion of studies that experimentally manipulate different categories of variation. For details on our study selection and classification procedures, see Appendix A.

In Panel A of Figure 1, we break down the studies by the method used. Although both methods are applied to most questions, notable differences appear across topics. The most popular areas, such as decision-making and the implementation of strategies in corporate settings, show a relative balance between simulations and human subject experiments. Other topics, including competition, environmental adaptation, and networks, are more often studied through simulations. We speculate that this pattern reflects structural constraints: it is difficult to recreate these dynamics with human subjects, and even more challenging to manipulate environments in ways that align with theoretical models. In contrast, areas like strategic human capital and entrepreneurship are more commonly studied through experiments—likely because they involve human behaviors that are harder to model computationally, and because they aim for empirical, policy-relevant estimates that simulations alone cannot provide.

In both approaches, scholars rely on controlled variation to link actions to outcomes. We reviewed each study to identify which dimensions were experimentally manipulated. From this analysis, we inductively developed a taxonomy that distinguishes between manipulations to the agent and changes to the environment (see Appendix A). We identify three sources of agentic variation (objectives, capabilities, and decision processes) and six sources of environmental variation (population, time horizon, choice landscape, payoff structure, information set, and spatial structure). Each study was coded according to this framework. Overall, we find that computational simulations make use of significantly more variation across both agentic and environmental dimensions. On average, a paper developing an ABM manipulates 3.7 categories, while human experiments vary only 2.1 of them. This pattern supports the view that simulation methods get closer to offering a form of "perfect control" (Harrison et al., 2007; Knudsen, 2024).

Panel B of Figure 1 reveals additional heterogeneity across methods. Simulations are more likely to introduce changes to the payoff structure, time horizon, agent population, and agent characteristics (Davis et al., 2007). Interestingly, some categories appear less amenable to computational manipulation. This may reflect the mechanistic nature of ABMs, which makes variations in elements such as agent objectives or information sets less theoretically interesting. Human experiments, by contrast, face stricter constraints on what can be manipulated, particularly in the features of the environment or the composition of the actor population. Yet because real-world observations are essential for testing causal relationships, these limited degrees of freedom pose a challenge when assessing the external validity of insights generated by more flexible ABMs. This difficulty helps explain why experiments are often confined to testing simple relationships, while empirical tests of ABMs remain rare.[3] These limitations suggest that strategy research could benefit from alternative tools that combine elements of human realism with some of the control afforded by computer simulations.

2.2 The Birth of AI Simulators

Recent advancements in generative AI have opened up many applications, particularly with the emergence of LLMs. These models are built on deep learning transformer architectures and trained on extensive text corpora, enabling them to generate human-like text in response to prompts (Bommasani et al., 2021). LLMs

[3] In the few cases where such experimental translation has been attempted, predictions from ABMs have shown mixed results, especially for widely used frameworks such as NK models (Billinger et al., 2014; Ganco, 2017).

have gained popularity due to their impressive performance across domains (Bubeck et al., 2023), in some cases surpassing human-level capabilities (Srivastava et al., 2022). Increasingly, LLMs are being adopted as tools in research (Grimes et al., 2023; Carlson and Burbano, 2025). As shown in Appendix Table B, most applications involve supporting research indirectly—through tasks such as dataset cleaning, scientific writing, and refining research instruments. However, LLMs are also beginning to play a more direct role in research, including generating hypotheses (Manning et al., 2024; Si et al., 2024).[4]

We anticipate that a central role of LLMs in strategy research will be as synthetic subjects (Bail, 2024; Grossmann et al., 2023; Park et al., 2023). This expectation is grounded in the growing evidence showing that LLM-powered agents exhibit strikingly human-like traits. When evaluated with personality tests and behavioral games, their responses are statistically indistinguishable from those of randomly selected human participants (Mei et al., 2024). Additional studies reveal that LLMs lean on heuristics similar to those used by humans and reproduce their cognitive biases, reasoning errors, and moral intuitions (Lampinen et al., 2024; Hagendorff, 2024). Horton (2023) demonstrates LLMs' strategic acuity by placing them in classic interactions such as the prisoner's dilemma and the dictator game. This methodology is gaining traction in computational social science, where LLMs serve as generative AI simulators that produce "silicon samples," or fully synthetic datasets, to study specific questions (Dillion et al., 2023; Sarstedt et al., 2024).

Despite their promise, recent studies also highlight several limitations of AI simulators (Aher et al., 2023; Gao et al., 2025). First, many language models have been trained on earlier research, raising the possibility that their outputs reflect memorized content rather than fresh reasoning. This is particularly problematic when directly prompting an LLM, for instance, to simply predict the outcome of prospective studies, although recent evidence does seem to downplay this concern (Luo et al., 2025). Second, while LLMs avoid some ethical constraints of human subjects research, they introduce new concerns, such as the spontaneous emergence of deceptive behavior in AI agents (Hagendorff, 2024). Finally, and perhaps most concerningly, it remains uncertain whether LLMs reliably mirror human behavior. Although fidelity may vary by task, some discrepancies seem systematic (Dillion et al., 2023; Jones and Steinhardt, 2022). Determining when an LLM provides a credible proxy for human responses is, therefore, an ongoing research challenge (Broska et al., 2024).

[4]Our focus is on how LLMs can be used in research and, more generally, strategizing endeavors. However, it must be noted that a florid literature is exploring the direct impact of AI on firm activities and performance (Csaszar et al., 2024; Dell'Acqua et al., 2023; Doshi et al., 2024; Jia et al., 2024). Relatedly, another strand of research studies the extent to which generative AI can replace humans in carrying out a variety of tasks, and the related labor-market implications (Eloundou et al., 2024; Felten et al., 2023). Our paper complements these strands of literature by focusing on the use of LLMs to develop strategy theory.

2.3 AI Agents as the Model Organisms for Strategy

Given ongoing concerns about whether LLMs can truly mimic human behavior, it remains an open question whether they are suitable for strategy research. A useful starting point is to draw a parallel with other fields that study highly complex phenomena. One of the clearest examples is biomedicine, which grapples with one of the most intricate and interactive systems we know: the human organism (Dawkins, 1976).[5] The use of synthetic subjects to build parsimonious theories about complex biological interactions is well established in this field. Scientists have long recognized that they can simulate the human body by studying *model organisms*, species that are biologically similar enough that insights gained may still plausibly apply to humans.

Model organisms, such as the lab mouse, are widely considered sufficiently analogous to humans to support the development of understanding about human biology (LaFollette and Shanks, 1995). Their primary value lies in theory generation: by observing how non-human animals respond to stimuli under controlled conditions, researchers can form empirical expectations about related human processes. These organisms complement other tools in biomedical theory development, including in vitro studies and clinical trials. In vitro studies isolate cells outside their natural context and offer high internal validity at low cost, but sacrifice realism. Clinical trials with humans provide the most realistic test of a treatment, but involve high costs and ethical risks. Model organisms serve as a crucial middle ground, balancing control with realism, even if it is intuitively clear that species like mice differ from humans in important ways (LaFollette and Shanks, 1995).

We argue that large language models (LLMs) may represent the first credible model organisms for strategy research, playing a role similar to animal models in biomedical science. LLMs share key features with strategic actors: they respond to incentives, adapt to context, and engage in dynamic, interactive behavior (Horton, 2023). Like lab mice in biology, they offer a flawed but valuable approximation for theory development (Tjuatja et al., 2023). Their strength lies in their ability to reflect complexity and generate unexpected variation, both of which are crucial for theory building (Tranchero, 2025). Once a researcher has a thesis and a basic experimental design, they can simulate it using LLMs, assess alignment with their intuitions, and iteratively adjust constructs, relax assumptions, or explore alternative conditions. This approach mirrors traditional computational experiments but trades some control for greater realism, offering a new tool to explore how strategic behavior unfolds in complex settings.

Yet caution is needed when using LLM-based agents in strategy research. The black-box nature of LLMs

[5]Further underscoring the parallel, it is no coincidence that the NK simulation approach—so widely adopted in strategy—was originally developed by Kauffman (1993) in evolutionary biology.

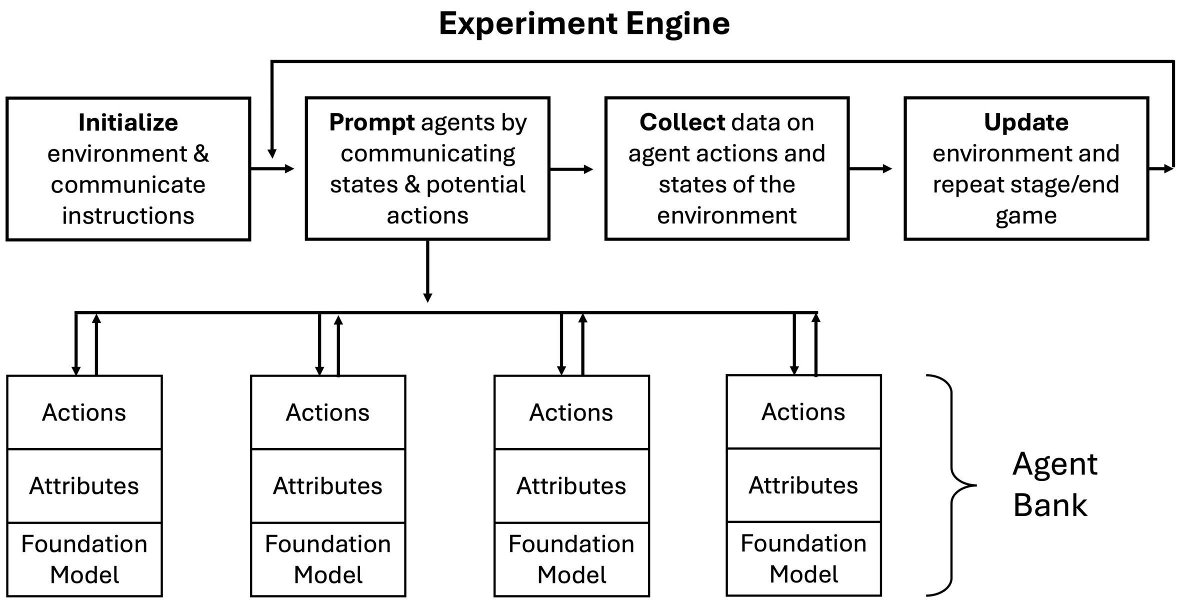
makes it more difficult to trace their behavior back to clear mechanisms compared to ABMs. And while their outputs often resemble human behavior, they do not constitute empirical evidence capable of falsifying theories. Still, we maintain that LLMs can play an important role in theory development, but in a manner guided by the researcher. From shaping the initial framework to interpreting results, synthetic subjects require a human "instigator" to steer inquiry and apply judgment (Agrawal et al., 2019; Felin and Holweg, 2024). This contrasts with efforts to fully automate theory generation through AI alone (Manning et al., 2024). In our view, experiments with AI agents are best understood as complementary tools for generating exploratory insights—insights that must ultimately be sharpened through more parsimonious ABMs and validated with human-subject experiments.

3 A Framework for Simulating LLM-Powered Experiments

In this section, our goal is to provide a general framework that enables strategy scholars to simulate experiments where multiple AI agents interact in a rich strategic environment. Our approach is systematic and bottom-up, placing micro-level autonomous AI agents in controlled settings, where they are given opportunities to interact organically with each other. These interactions can produce emergent behaviors that are not presumed by the researcher from the outset and are unknown to the individual LLMs—much like when biomedical researchers observe lab mice after an experimental treatment. Throughout the simulations, each agent's decisions, rationales, and resulting outcomes are carefully recorded, enabling an in-depth study of strategic interactions.

Because these experiments are conducted entirely in silico, they require some basic computational infrastructure. The researcher must first set up a central engine that administers the experimental process virtually. This engine is responsible for creating AI agents, assigning their roles and behaviors, delivering instructions, updating information as the experiment unfolds, and recording outcomes. Figure 2 provides a stylized overview of this process. Researchers have flexibility in how to build the engine and can implement it in nearly any coding environment. One option is to use existing tools such as Expected Parrot Domain-Specific Language (EDSL), an open-source Python package that connects user prompts to the GPT-4 API (Horton and Horton, 2024). EDSL simplifies many of the tasks involved, including launching multiple AI agents and managing structured, sequential interactions. We make use of this functionality in the case study that follows. Technical details are provided in Appendix C. Because the core functions are agnostic to the specific experiment, the infrastructure is highly adaptable and requires only minor adjustments when applied to new studies. Once in place, the researcher can run a virtually unlimited number of simulations. Unlike setups that replicate survey-based experiments (Ashokkumar et al., 2024), this engine-based approach is well suited for studying complex, repeated interactions among multiple agents.

Figure 2: Stylized Framework to Run LLM-powered Strategy Experiments



Note: This figure presents a stylized depiction of the central deterministic engine that powers an interactive experiment. The engine spawns the AI agents (each consisting of an action space, set of attributes, and underlying foundation model) and then administers the experiment in silico following the steps above.

The researcher must then specify the key parameters of the experiment. Because the entire setup is a computer simulation, researchers can manipulate all of the degrees of freedom outlined in Panel B of Figure 1. This includes variation at the agent level, such as goals, capabilities, and decision-making rules, which can be implemented by prompting the LLMs with additional scripts. It also includes variation in the environment, such as the number of agents, time horizon, choice landscape, payoff structure, information set, and spatial features, all of which can be easily varied. In addition, the researcher can determine how the experiment is administered, including whether decisions are made sequentially or simultaneously and whether agents are allowed to communicate. They can also choose which outcomes to collect, whether quantitative data for econometric analysis or qualitative responses to support abductive theorizing. The main degrees of freedom available to the researcher are described in Appendix D. As with lab experiments and ABM simulations, these design choices should be grounded in the starting theory that motivates the research question (Davis et al., 2007; Ott and Hannah, 2024).

Finally, this framework opens new avenues for uncovering the causal mechanisms behind observed results. The simplest approach is to ask the model to explain its decisions with an explicit prompt. These explanations can be enriched through chain-of-thought prompting, which records the model's reasoning steps and reveals how decisions unfold (Wei et al., 2022). Although not yet common, researchers may soon be able to study how outcomes shift in response to direct changes to the foundation model. As the field of model interpretability advances, modifications to the model's architecture or activation patterns may become accessible (Lindsey et al., 2025). While LLMs may not always "know" exactly why they made a particular choice (not unlike humans) and their responses may involve some degree of hallucination, the researcher can assess the plausibility of these rationales. Ultimately, it remains the researcher's responsibility to interpret the results and draw theoretical insight from them.

4 Application: the Exploration-Exploitation Dilemma

4.1 A Theory of Exploration with Data Spillovers

We illustrate the use of AI agents as synthetic subjects through a canonical problem in strategy. In many innovation and entrepreneurship settings, actors must choose from a set of options with uncertain value. Typical examples include venture capital funds deciding which startups to back (Lerner and Nanda, 2020), or pharmaceutical firms selecting which genes to target for drug development (Tranchero, 2025). Some options may already have known value, in which case the actors must decide whether to exploit known options or explore new ones (March, 1991). Although this can be an individual decision (Levinthal, 1997), it can also depend on the actions of others (Lazer and Friedman, 2007), introducing strategic interdependence. For example, firms frequently learn from their competitors (Krieger, 2021), and must decide whether to build on existing innovations or to invest in R&D themselves, knowing that the results may later become public. What strategy should a manager pursue in such settings?

To explore this question, a recent paper studied an exploration–exploitation setting with data spillovers (Hoelzemann et al., 2025). The authors develop a closed-form model in which agents choose among risky projects whose value is revealed only through exploration. The paper examines how public data on projects' value affects firm behavior. Notably, the authors identify a parameter space where revealing the value of a medium-return project causes exploration to collapse due to free-riding behavior. The counterintuitive result is that public data makes all firms worse off than if no information had been shared. They test this prediction in a lab experiment, where undergraduate students play an incentivized game simulating exploration under uncertainty, and find empirical support. This work provides a compelling starting theory for studying what is often referred to as the *streetlight effect*, or the tendency to search where data is easiest to access, even when doing so leads to inferior outcomes (Haynes et al., 2018; Vuculescu, 2017).[6]

[6] This term comes from the parable of a drunkard who searches for his keys beneath a streetlamp because that is "where the light is", even though he knowingly lost them elsewhere.

We focus on this study for several reasons. First, the online lab experiment features strong strategic interdependencies and uncertainty. These elements are common in real-world settings, making it an ideal test case for evaluating whether LLM-based simulations can contribute meaningfully to strategy research.[7] Second, the experiment relies on a single set of assumptions and parameters, leaving room for theoretical elaboration. We take up that task by using LLM-based simulations to first replicate and then extend the original findings. Third, the study is grounded in a closed-form mathematical framework. This allows us to benchmark both human and LLM behavior against theoretical optima, providing a precise context for interpretation. Finally, we are confident that our simulations represent out-of-sample predictions. A common concern in studies that replicate human-subject experiments is that LLMs may simply reproduce results seen during training. In our case, we verified that GPT-4 is unfamiliar with the experiment, which reduces this risk.[8]

4.2 The Online Lab Experiment: Searching Mountains for Hidden Gems

The experiment of Hoelzemann et al. (2025) features a group of participants searching across a virtual range of mountains for hidden gems. The setup is shown in Panel A of Figure 3. There are $n = 5$ players and $m = 5$ mountains. Three of the mountains contain topazes, one mountain contains a ruby, and one mountain contains a diamond. While the dollar value of each gem varies by round, the diamond is always valued more than the ruby, which in turn is valued more than the topazes. The location of these gems is unknown from the outset, and players only know the number of gems and their values. The experiment consists of two periods: in the first period, each player selects a mountain in sequential order. After everyone has finished selecting, the gems behind the selected mountains will be revealed. In the second period, players choose again, this time with the knowledge of gems revealed in the first period. Each player earns the sum total of the payoffs from their choices in both periods. The payoffs are non-rival, which means there is no penalty for choosing the same mountain as other players.

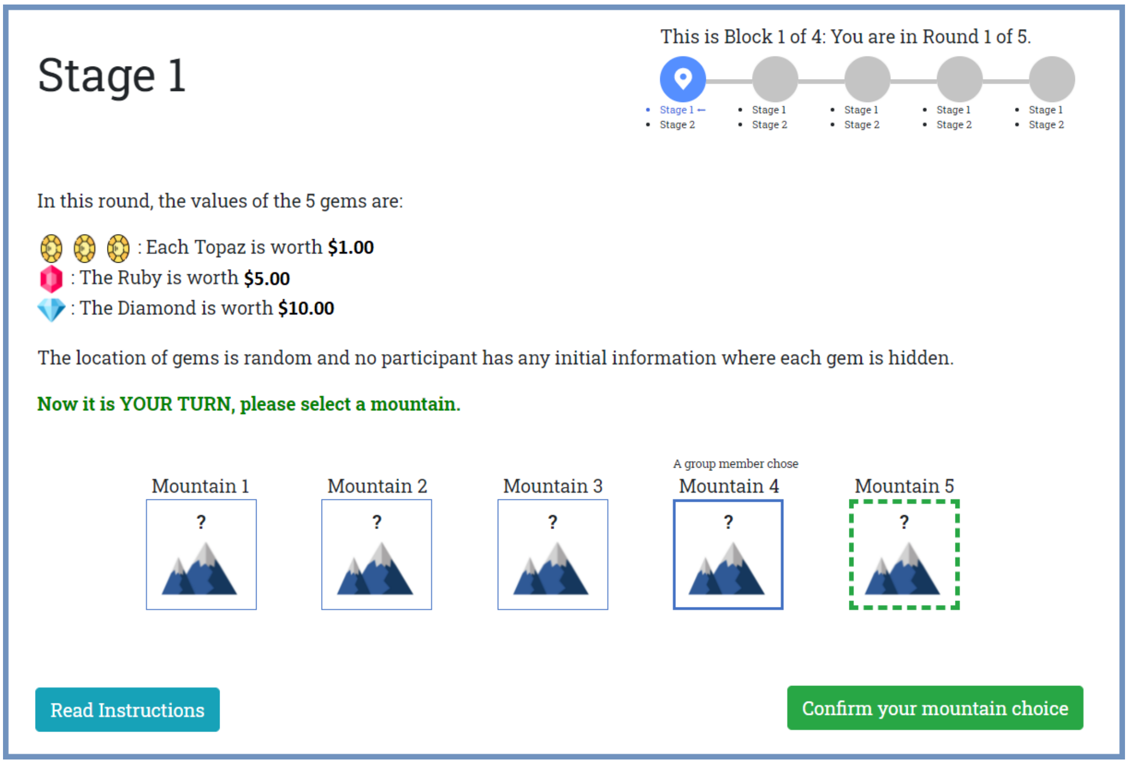
In the baseline condition, participants are not shown any data on the location of the gems. However, the experiment has several other treatment conditions, each providing partial initial data on payoffs. These are depicted in Panel B of Figure 3. In the low-value condition, the location of one topaz is revealed at the beginning of the experiment, while the same happens with the ruby in the medium-value condition and with the diamond in the high-value condition. This design allows researchers to observe how the availability (and

[7] An additional advantage is that we have full access to the original materials and results. In most replication studies, researchers must approximate the original design as closely as possible. Here, we can compare results directly and introduce extensions that remain faithful to the underlying model.

[8] We verified this by prompting GPT-4 to describe the experiment in Hoelzemann et al. (2025). The responses were inaccurate and largely hallucinated. This is consistent with the fact that the version of GPT-4 we used was last updated in December 2023, while the working paper was released in 2024.

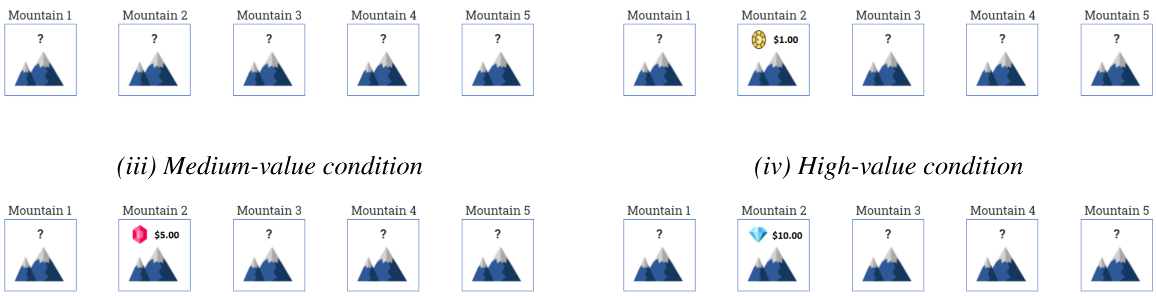
Figure 3: The Experimental Platform from Hoelzemann et al. (2025)

Panel A: User interface



Panel B: Examples of no-data condition and data conditions

(i) No-data condition *(ii) Low-value condition*



Note: This figure depicts the software platform used in the online lab experiment of Hoelzemann et al. (2025). Panel A shows how the interface is seen by participants in the no-data condition. The values of the gems for each round are shown in the upper left corner. In this example, the user can see that Mountain 4 has already been picked and decides to select Mountain 5. Panel B depicts the four experimental conditions. For instance, in plot (iii), the mountain uncovering the ruby is revealed at the start of the round. The figure is reproduced with permission from Hoelzemann et al. (2025).

nature) of initial payoff-relevant data influences exploration strategies and outcomes. The study involved 350 participants and was conducted over 1400 rounds, with each round consisting of 5 players and 2 periods.

We reproduce this experiment using our framework. The key parameters that we supply from our framework are shown below in Table 1. We keep the same setup as Hoelzemann et al. (2025). However, this time, it is a group of AI agents selecting which mountains to explore each round. Using our framework, we spawn the five AI agents at a time, familiarize them with the instructions of the experiment, and prompt them sequentially for their choice of mountains.[9] Once an AI agent chooses a mountain, the code updates the conditions of the environment and informs the other AI agents about this choice (see Appendix C for an excerpt of the script we use). In total, we simulate 500 rounds of the experiment.

Table 1: Application of the Framework to Run LLM-powered Strategy Experiments to Hoelzemann et al. (2025)

Parameter	Description
Environment	The engine keeps track of the game state, including which mountains have been explored, which gems lie underneath, and which agents have picked which mountain. *Population:* There are 5 AI Agents. *Time Horizon:* There are 2 periods. *Choice Landscape:* There are 5 choices (i.e. mountains). *Payoff Structure:* Each mountain is associated with one of 3 payoffs, depending on whether it holds a topaz, ruby, or diamond. *Information Set:* Agents know the value and distribution of gems, but they do not know the locations. *Spatial Features:* There are no spatial features to the environment.
Agent	*Objectives:* Each agent is explicitly instructed to maximize their individual earnings. *Capabilities:* Each agent can pick one mountain at a time. There is no private knowledge. *Decision-rules:* No decision-rules or algorithms for choosing mountains are specified.
Turn-taking procedures	The experiment uses sequential decision-making: each agent is allotted a "turn" to make their decision. This provides an implicit mechanism for agents to coordinate their exploration.
Communication channels	AI agents are not allowed to verbally communicate with each other during the course of the experiment.
Outcomes	We record the numerical earnings of AI agents, whether or not each group made a "breakthrough" and discovered a diamond, and the number of mountains explored. We also solicit AI agents for qualitative explanations behind their decisions.
Interventions	We make isolated changes to the information set. In particular, we have different initial data conditions. In the "no data" condition, the experiment begins with five unknown mountains, while in the other conditions, it begins with one mountain revealed as either high, low or, medium value.

Note: This table illustrates how we specify the key parameters of our framework to replicate the online lab experiment of Hoelzemann et al. (2025). See the text and Appendix D for more details.

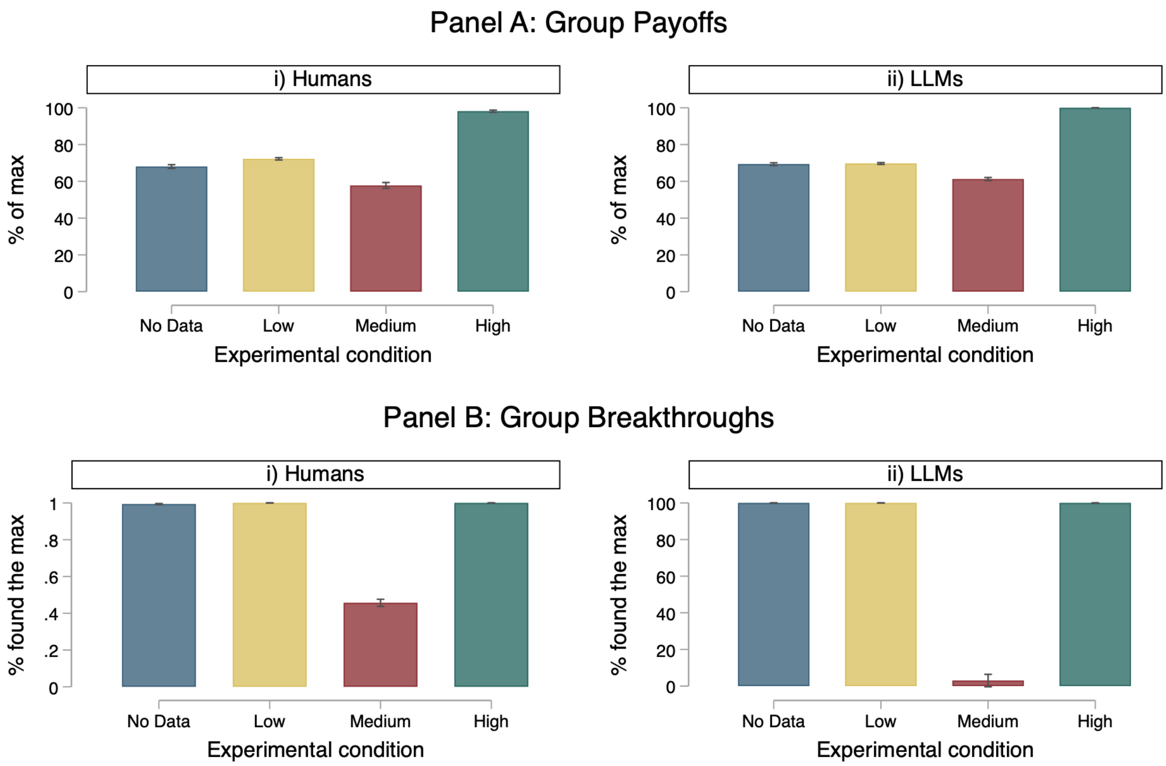
4.3 Comparing Human Subjects and AI Agents

We begin by replicating the original experiment using AI agents. The primary outcome of interest is the group payoff, defined as the total sum of all individual payoffs in a group for a given round. We plot the mean group payoff across experimental conditions, along with 95% confidence intervals (Panel A of Figure 4). The original results from the online lab experiment appear in the upper left corner (plot i). When participants are shown the location of the medium-valued gem, they tend to earn the least. Further analysis shows that participants avoid the risk of searching for the higher-valued diamond and instead herd on the safer ruby (Appendix Figure E1). This choice boosts payoffs in the early stages but ultimately leads to lower total

[9] We describe EDSL and release open access code to replicate and extend our analysis in Appendix C. EDSL also enables us to prime the agents to have different objectives or features. For more details about this, see Section 5.3.

earnings by the end of the round (Appendix Table E1). In contrast, when no information is provided or only the location of a low-value gem is revealed, participants are more likely to continue searching, increasing their chances of uncovering the diamond.

Figure 4: Comparing the Lab Experiments with Human Subjects and AI Agents



Note: This figure compares the results of the lab experiment when using AI agents versus human subjects. In Panel A, the outcome of interest is the average group earnings (as a percentage of the maximum possible earnings). In Panel B, the outcome of interest is the mean likelihood of a breakthrough, which occurs when at least one group member finds the diamond. In each panel, plot i) shows the results obtained in the original lab experiment using human subjects, while plot ii) shows the results obtained from the simulated experiments using AI agents.

We examine how closely AI agents replicate this strategic behavior in plot ii). The results are strikingly similar, even in this setting with complex interdependence. AI agents earn the least when shown the location of the ruby. Like human participants, they choose not to explore further and instead converge on the safer option (see Appendix Figure E2 and Appendix Table E2). When shown only the location of the topaz, or when no information is provided, the agents effectively coordinate their search, using the sequential order of choices to explore all options and ultimately achieve higher payoffs. When the diamond's location is revealed, they immediately select it to maximize earnings. In short, LLMs reproduce the same behavioral patterns of human participants and capture the key insight of Hoelzemann et al. (2025): more data is not necessarily beneficial in search contexts.

Our second outcome of interest is whether the group achieves a breakthrough (Panel B), defined as at least one member discovering a diamond. Once again, the patterns are broadly similar. Both AI agents and human participants achieve a breakthrough in nearly every round where the ruby is not revealed, either because they search and find the diamond or because it is revealed to them. However, a key difference emerges when the ruby is shown. In this case, humans achieve a breakthrough in roughly 45% of rounds, while AI agents almost never do. As Hoelzemann et al. (2025) shows, breakthroughs should not occur under rational behavior, suggesting that LLMs adhere more closely to theoretical predictions. This finding aligns with other research showing that language models tend to behave more rationally than humans (Hagendorff, 2024), revealing a key limitation. Qualitative interviews from the original experiment indicate that some humans deviated from the ruby out of boredom or randomness, both behaviors that appear largely absent among AI agents. Even so, the overall behavioral patterns across humans and AI agents remain strikingly similar.

One might worry that the similarity in outcomes is merely coincidental. It is possible, for example, that LLMs arrive at their decisions through processes unrelated to the strategic reasoning that guides human behavior. To probe this further, we asked each LLM to explain its choices. These responses were highly revealing. As one representative LLM put it:

> *I choose Mountain 5 because it is revealed to contain a Ruby, which is valued at $6. The certainty of obtaining a Ruby outweighs the risk of potentially finding a Topaz, despite the possibility of finding a Diamond worth $10 in another mountain.*

This suggests that LLMs follow the same behavioral logic as human decision-makers. While we cannot rule out the possibility of hallucination or post-hoc rationalization, the simplest and most plausible explanation is that their choices reflect the stated reasoning—especially given the alignment between their rationales, decisions, and outcomes, and the patterns observed in human participants. Further supporting this interpretation, we reran the experiment using reasoning models that articulate their thought process in real time. These transcripts reveal the same logic unfolding step by step as the model works through the problem.[10] Taken together, this replication serves as a first step in assessing how realistically LLMs can simulate human

[10]The following is a transcript of one reasoning LLM's internal thought process during the task:

> ****Evaluating mountain choices****: I need to choose which mountain to pick based on what I've learned. Mountain 1 is revealed to have a Ruby, valued at 8. Since there's only one Ruby, Mountains 2-5 must have the remaining values: one Diamond worth 12 and three Topazes worth 3 each. To maximize the expected value, I should note that picking Mountain 1 guarantees an 8, while the other mountains have a lower expected value of 5.25. Choosing Mountain 1 seems like the best option!
>
> ****Deciding on the best choice****: I need to calculate the expected payoff from picking an unknown mountain. By doing the math, I find that 0.25 times 12 equals 3, and 0.75 times 3 equals 2.25, which totals 5.25. This is clearly less than the guaranteed value of 8 from Mountain 1. Therefore, my best rational choice is to select Mountain 1 with the Ruby. I'll respond simply with "Mountain 1" and provide a comment to explain my reasoning.

behavior in strategy experiments. By analogy to a machine learning problem, if AI simulators can replicate human responses on held-out data with this level of fidelity, it offers a prima facie case for using them to simulate new, untested experimental designs. Moreover, the quote illustrates how simulated experiments using AI agents can help with explaining patterns in an abductive manner—a fundamental mode of theorizing in strategy research (Makadok et al., 2018).

5 Extending the Starting Theory using LLMs

We now take advantage of the substantial control afforded by the computational setup—both over the agent and the environment—as well as the low cost and speed of running simulations, to explore a series of extensions to the starting theory. Specifically, we a) vary the experimental parameters, b) relax key theoretical assumptions, and c) introduce heterogeneity in agent preferences. These extensions allow us to probe the boundary conditions of the original experiment and uncover new potential mechanisms. A summary of the full set of extensions and their results is presented in Table 2.

5.1 Varying the Experimental Parameters

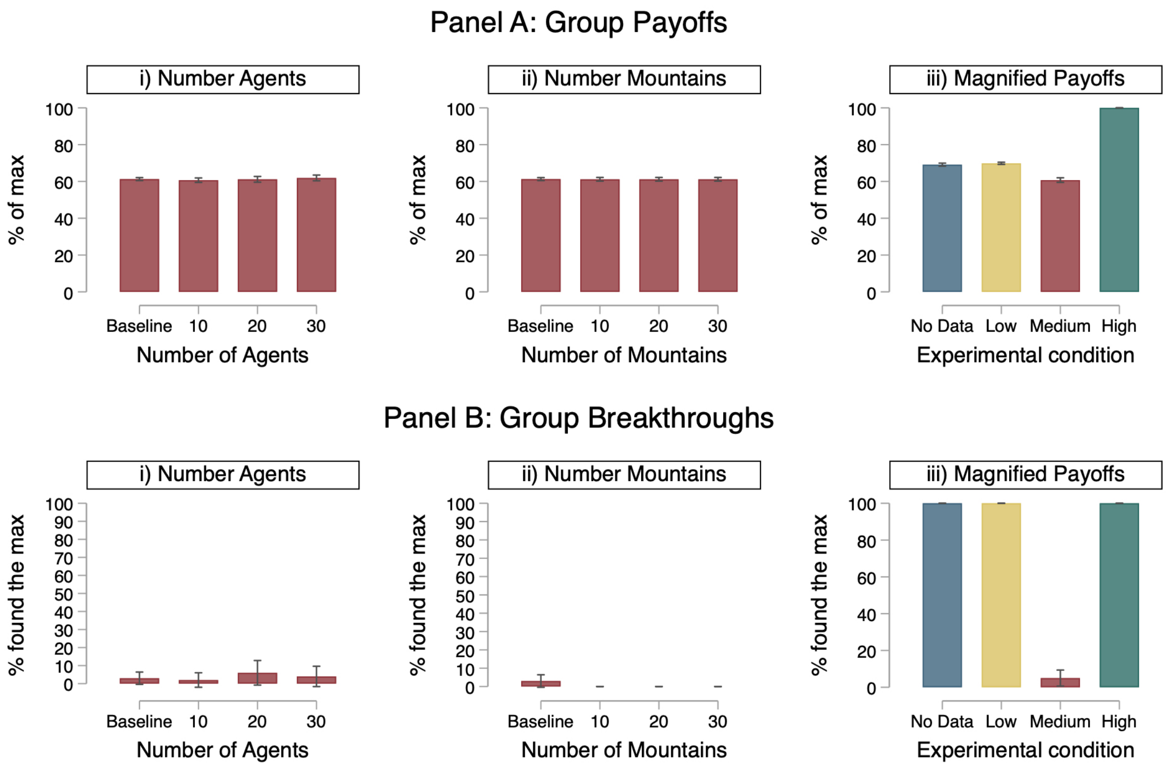
The first set of extensions focuses on the experimental setup itself. When researchers test a theory through an experiment, costs usually limit them to trying a fixed number of parameters. Some of these choices are likely innocuous, but others may influence the results in unforeseen ways. To investigate this in the context of Hoelzemann et al. (2025), we conduct 800 additional rounds of AI-based simulations, varying one of three major dimensions at a time.

A. Varying Population: We begin by varying the number of agents while keeping the number of mountains fixed. To simulate more crowded search environments, we rerun the original experiment with larger groups of AI agents. As shown in plots i) of Figure 5, increasing the number of participants by increments of ten does not lead to greater exploration. Even in larger groups, AI agents continue to cluster around the ruby and often fail to make a breakthrough. When asked to explain this behavior, the agents point to a dynamic we had not previously considered, revealing a new and intriguing theoretical mechanism: social conformity. As one AI agent explained:

> *"Given that 19 agents have already chosen Mountain 1 and each will receive the full value of the gem, it indicates a common strategy to secure a known and relatively high value".*

The explanation above mirrors earlier findings that emulation can shape behavior in group settings (Lazer and Friedman, 2007; Puranam and Swamy, 2016), but was completely absent from the original study of Hoelzemann et al. (2025). This highlights how LLM-based simulations can surface meaningful theoretical tensions worth exploring further.

Figure 5: Changing the Experimental Parameters in LLM-based Experiments



Note: This figure shows how the outcomes of the original experiment change when we adjust the experimental parameters using LLMs. In Panel A, the outcome of interest is the average group earnings (as a percentage of the maximum possible earnings). In Panel B, the outcome of interest is the mean likelihood of a breakthrough, which occurs when at least one group member finds the diamond. In plots i), we increase the number of agents from the baseline (5 agents) by up to 30 agents, holding the number of mountains fixed. We only focus on the medium-value data condition. In plots ii), we increase the number of mountains from the baseline (5 mountains) by up to 30 mountains, holding the number of agents fixed. We once more focus on the medium-value data condition. In plots iii), we increase the dollar values of gems a millionfold. Here, we look at all four data conditions.

B. Varying Choice Landscape: Next, we vary the number of mountains while keeping the number of agents fixed. In this setting, coordinated search no longer guarantees a breakthrough, reflecting larger search spaces where not every option can be explored. We rerun the original experiment with a progressively increasing number of mountains. As shown in plots ii) of Figure 5, we again find little change in behavior: AI agents continue to cluster around the medium-value option. They are not swayed by the greater absolute number of diamonds or rubies and follow the same decision logic as before. If anything, exploration declines, suggesting that larger search spaces may further reduce experimentation. One potential reason behind this result is that some agents appear to interpret the herd behavior of others as a signal that they have missed valuable information (Banerjee, 1992):

> *"The fact that many other participants have also chosen Mountain 2 reinforces the idea that it is a preferable choice given the available information"*.

This interpretation has some precedent in the strategy literature, where firms often struggle to evaluate complex opportunities and rely on imperfect heuristics (Dahlander and Piezunka, 2014) or mimetic isomorphism (Haveman, 1993) to deal with uncertainty. Whether similar dynamics would emerge when human subjects face a larger choice set is a theoretically compelling question that can be tested in future laboratory studies.

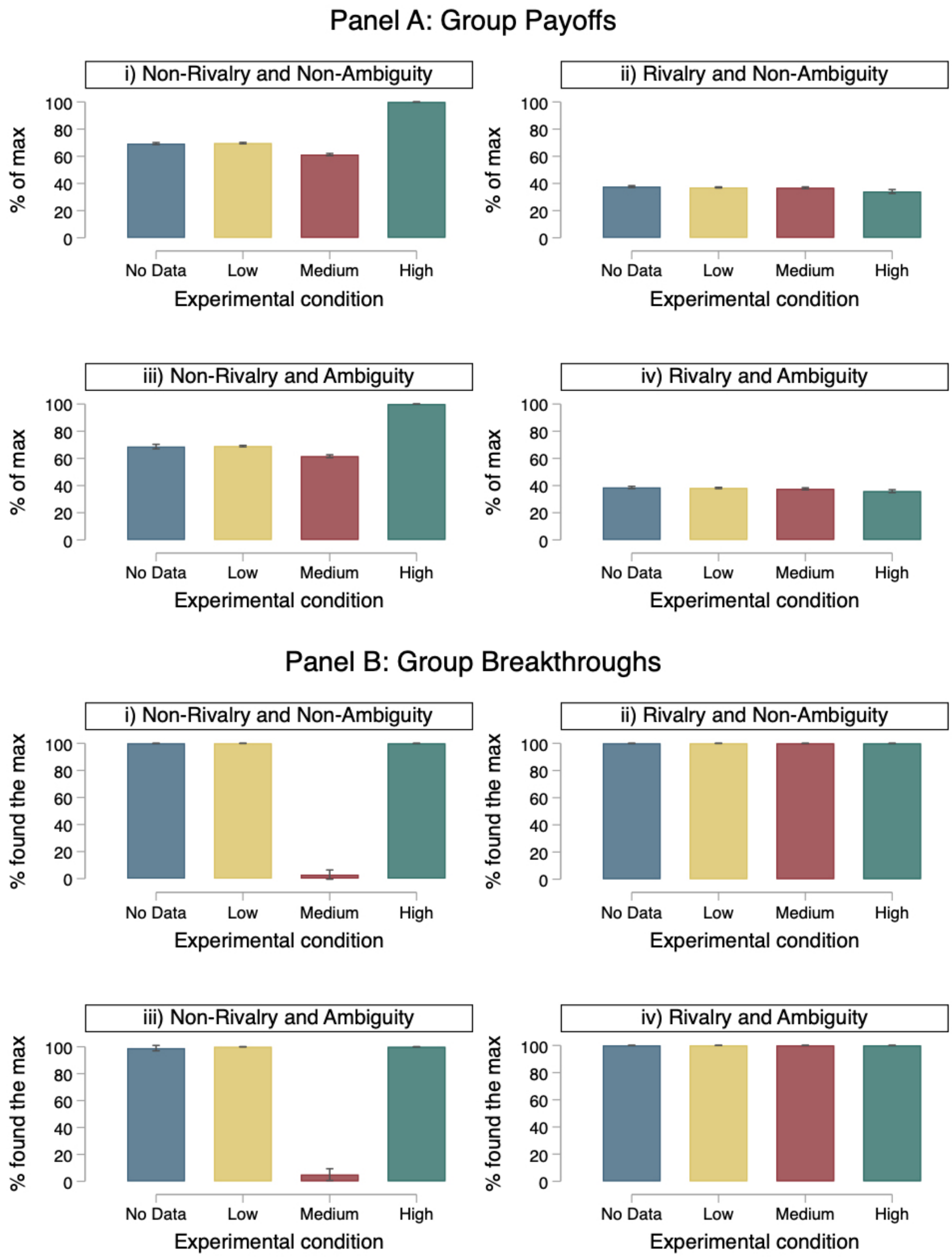
C. Varying Payoff Structure: The final experimental feature we vary is the absolute magnitude of the payoffs. While the theoretical framework in Hoelzemann et al. (2025) defines the relative values of the gems, the authors had to select specific dollar amounts for the lab experiment. They chose amounts below 15$ per gem to keep the experiment affordable, but we might wonder how the decisions would change if players were earning up to several millions in a given round. Of course, such an experiment would be infeasible with human subjects, who would be entitled to the payouts, but it is easily simulated with AI agents. We rerun the original experiment with payoffs scaled up by a factor of one million. The results, shown in plots iii) of Figure 5, mirror the baseline. This aligns with the theoretical prediction that only the relative value of each gem should influence decisions. Whether this holds for humans is less clear, as the psychological effects of extreme stakes may distort reasoning. Still, the ability to simulate such conditions with AI agents offers a valuable base expectation that would otherwise be unattainable.

5.2 Relaxing Theoretical Assumptions

The next set of extensions that we explore focuses on the underlying theoretical framework. Two key assumptions are embedded in the original model: that payoffs are non-rivalrous and non-ambiguous. Non-rivalrous payoffs imply that multiple agents selecting the same gem do not have to share the reward, while non-ambiguous payoffs assume that the distribution of gem values is known by agents from the outset. These assumptions also guided the design of the original online lab experiments. Understanding how sensitive the results are to these assumptions is important for evaluating the robustness of the theory. In particular, it is useful to know whether free riding persists when payoffs are rivalrous or when information is incomplete. To address this, we conduct an additional 1,500 rounds of simulated experiments, relaxing each assumption individually. The baseline case, where payoffs remain non-rivalrous and non-ambiguous, is shown in plots i) of Panels A and B in Figure 6.

We begin by changing the payoff structure to introduce rivalry, meaning that agents who choose the same

Figure 6: Relaxing Theoretical Assumptions in LLM-based Simulations



Note: This figure shows how the outcomes of the original experiment change when we relax key theoretical assumptions using LLMs. In Panel A, the outcome of interest is the average group earnings (as a percentage of the maximum possible earnings). In Panel B, the outcome of interest is the mean likelihood of a breakthrough, which occurs when at least one group member finds the diamond. Plots i) show the baseline results, where there is no rivalry or ambiguity in payoffs. In plots ii), we introduce rivalry in payoffs, which means agents who choose the same gem do not have to split the pay. In plots iii), we introduce ambiguity in payoffs, which means the distribution of gems is known from the outset. In plots iv), we introduce both rivalry and ambiguity in payoffs.

gem must divide the reward. The results are shown in plots ii) of Panel A and Panel B. Rivalry reverses the original pattern of findings and reveals a potentially important boundary condition for the theory. Agents no longer earn the least (Panel A), nor do they achieve fewer breakthroughs (Panel B), when the medium-value gem is revealed. Because they can observe others' choices in sequential order, agents often select unchosen mountains to avoid splitting payoffs. Qualitative responses from AI agents support this behavior and show how competition acts as a strong incentive for exploration. As one agent put it:

> *I chose Mountain 5 to avoid competition and potentially discover the Diamond or another Topaz, maximizing my potential earnings by not sharing the already known Ruby value in Mountain 3 with another agent.*

As a result, payoffs begin to converge across the four conditions, though at a significantly lower level. This mirrors the effect of competition eroding profits in a market.

Next, we manipulate agents' information set by withholding information about the number of gems of each type. We begin by keeping payoffs ambiguous but non-rivalrous. The results of this variation are shown in plots iii) of Panel A and Panel B. Exploration continues to collapse under these conditions. Group payoffs (Panel A) and breakthroughs (Panel B) closely mirror the baseline: AI agents earn the least and rarely achieve a breakthrough when they know the location of the run-of-the-mill outcome. This suggests that the theory may not depend on prior knowledge of the full payoff distribution, but rather on knowledge of the relative value of available options.

Finally, we examine the case where payoffs are both rivalrous and ambiguous. This allows us to assess which assumption has a stronger influence on agent behavior, and whether relaxing both leads to any interaction effects. The results are presented in plots iv) of Panels A and B. Herding disappears entirely under these conditions, and agents behave almost exactly as they do when payoffs are rival but information is complete. This suggests that rivalry plays a dominant role in shaping behavior and may be the critical condition driving exploratory dynamics. This insight reveals a boundary condition that would have been difficult to anticipate without using LLMs as synthetic subjects.

5.3 Manipulating Agent Objectives

The final set of extensions focuses on the agents themselves. Hoelzemann et al. (2025) suggest that risk aversion and non-pecuniary objectives may help explain the experimental results and potentially influence the theoretical predictions. Studying such factors in human experiments is difficult, as varying participants' risk tolerance or motivations can be impossible. With AI agents, however, this becomes much more feasible. A key feature of LLMs is their capacity to adopt specific preferences, views, demographics, or personalities through prompt conditioning (Horton, 2023; Aher et al., 2023). This simply involves priming the model with tailored scripts. To test this in practice, we conduct 1,000 additional rounds of simulated experiments in which the LLMs are endowed with different preference structures.

We begin by introducing higher risk tolerance. We rerun the original experiment with increasing numbers of risk-loving agents, starting with one and adding one at a time until all five agents exhibit risk-seeking behavior.[11] We focus on the condition in which the medium-value gem is revealed. More breakthroughs occur (Figure 7, Panel A) as a result of increased exploration driven by risk-loving agents (Figure 7, Panel B). To verify that the priming has the intended effect, we review the agents' rationales, which provide useful insight. As one risk-seeking agent explains:

> *"Given my risk-loving nature and the information available, I choose Mountain 1. This decision is based on the fact that the Ruby has already been found in Mountain 4, and another agent has chosen Mountain 5. Since the Diamond has not yet been discovered and I prefer taking risks for potentially higher rewards, I opt for Mountain 1, which has not been chosen by any other participant and still holds the possibility of containing the Diamond."*

Interestingly, as more agents are primed to take risks, the remaining unprimed agents also begin to explore more. This suggests that risk tolerance can initiate an implicit coordination process. As one non-risk-loving agent tells us, exploring becomes more attractive when the likelihood of collectively uncovering the diamond increases:

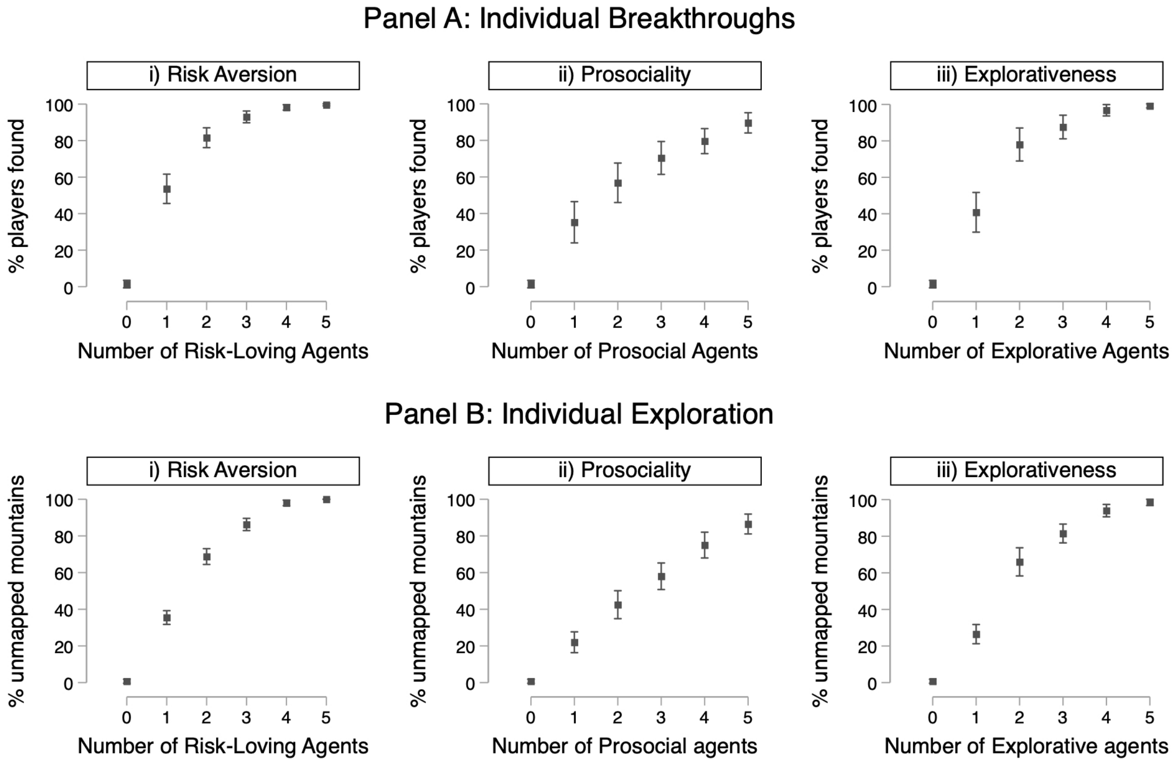
> *"I choose Mountain 1 because Mountains 3, 4, and 5 have already been chosen by other participants. Given that Mountain 2 is the only other remaining option aside from Mountain 1, I randomly select Mountain 1.*

This implies that only a subset of agents needs to be risk-loving to generate sufficient exploration, a prediction that can be tested in future laboratory experiments.

Next, we introduce motives beyond profit maximization, beginning with pro-social behavior. We rerun the

[11] Before running this extension, we assessed the baseline risk propensity of AI agents using the traditional Holt and Laury task used in behavioral economics. Consistent with previous findings, we confirm that unprimed AI agents tend to be risk-neutral, similar to the average participant in academic lab studies (Mei et al., 2024). We then repeat the preference elicitation task after priming agents with their respective risk profiles, confirming that the priming significantly alters responses in the expected direction.

Figure 7: Incorporating Agent Preferences in LLM-based Simulations



Note: This figure shows how the outcomes of the original experiment change when we incorporate heterogeneous agent preferences using LLMs. In Panel A, the outcome of interest is the share of agents that achieve a breakthrough in the second period.. In Panel B, the share of agents that chose an unmapped mountain in the first period. In plots i), we add risk-loving agents, starting with one risk-lover until all five agents are risk-loving. In plots ii), we add pro-social agents, and in plots iii) we add agents with a taste for exploration. For all three analyses, we focus on the medium-value data condition.

original experiment with varying numbers of pro-social agents, starting with one out of five and incrementing until all are pro-social. These agents are instructed that their success depends on whether others find the diamond, which encourages them to generate information that maximizes the group's chances of success. We again focus on the condition where the middle-value gem is revealed. The results, shown in plots ii) of Panel A and Panel B in Figure 7, show that exploration increases steadily with the number of pro-social agents, as does the rate of breakthrough. The agents' rationales offer valuable insight:

> *"Since I am pro-social and aim to give other participants the chance to discover a diamond, I will choose a mountain that has not been selected by others yet. All other agents have chosen Mountain 4, which contains a ruby. To maximize the group's potential earnings and to explore the possibility of finding a diamond, I will choose Mountain 1.*

Finally, we introduce an explicit preference for exploration. We rerun the original experiment with varying numbers of explorative agents, starting with one and increasing until all agents are primed to prioritize exploration. These agents are instructed that their sole objective is to find the diamond and that they should continue searching until they succeed. The results, shown in plots iii) of Panel A and Panel B in Figure 7, closely resemble those observed with risk-loving agents. As expected, exploration increases steadily with the number of explorative agents.

Table 2: Summary of Results Using LLMs as Synthetic Subjects

Intervention	**Description**	**Rounds (#)**	**Cost ($)**	**Hypothetical Payment ($)**	**Summary of findings**
Baseline	The original experiment with 5 players and 5 mountains	500	250	50,000	Players herd around the suboptimal option, lowering payoffs
Extension 1: Varying the experimental parameters					
Varying group size	We vary the size of groups from 5 agents to 30 agents	150	300	60,000	No change in exploration: social conformity reinforces the decision to herd around suboptimal option
Varying choice landscape	We vary the number of mountains from 5 mountains to 30 mountains	150	75	15,000	No change in exploration: agents unswayed by greater number of absolute options
Varying payoff magnitude	We amplify gem values one millionfold	500	250	-	No change in exploration: theory depends only on relative magnitudes of payoffs
Extension 2: Relaxing theoretical assumptions					
Relaxing payoff rivalry	Agents who choose the same gem must split the payoffs	500	250	50,000	Rivalry breaks the herding effect: agents want to avoid splitting payoffs
Relaxing payoff ambiguity	We no longer inform the agents how many gems of each kind there are.	500	250	50,000	No change in exploration: theory does not depend on prior information about distribution
Relaxing payoff rivalry and ambiguity	Agents must split payoffs, and we no longer inform agents of the distribution	500	250	50,000	Rivalry exhibits a greater force than ambiguity, continuing to break the herding effect
Extension 3: Manipulating agent preferences					
Incorporating risk-aversion	Some portion of the group is now risk-loving, varying from 1 agent to all 5 agents.	500	250	50,000	Risk aversion breaks the herding effect - exploration increases with the number of risk-lovers
Incorporating pro-sociality	Some portion of the group is now pro-social, varying from 1 agent to all 5 agents.	250	125	25,000	Pro-sociality breaks the herding effect - exploration increases with the number of pro-social agents
Incorporating explorativeness	Some portion of the group is now explorative, varying from 1 agent to all 5 agents.	250	125	25,000	Explorativeness breaks the herding effect - exploration increases with the number of explorative agents
Total		**3800**	**2125**	**$375,000**	*(Excludes payoffs from amplified case)*

Note: This table provides an overview of all the results from this study. Column 1 lists the extensions we introduce. Column 2 provides a description of each extension. Column 3 indicates the number of rounds simulated. Column 4 shows the approximate total cost in GPT-4 credits as of July 2024. Column 5 highlights the amount we would have needed to pay to human subjects. Column 6 summarizes the key result. Row 2 covers the baseline replication of the original lab experiment. Rows 4-6 cover changes in the experimental setup. Rows 8-10 cover key theoretical assumptions being relaxed. Rows 12-14 cover heterogeneous agent preferences being introduced.

5.4 Comparing Simulations to Direct Elicitation

So far, we have used LLMs to run experiments using AI agents as synthetic subjects, applying a bottom-up approach that parallels ABMs. However, another possibility is to use LLMs to predict the outcomes of experiments directly. Large language models have become increasingly capable of handling complex tasks, and their performance continues to improve. In this setting, an LLM could generate outcome predictions along with explanations for its reasoning. These justifications could help verify the results and shed light on underlying mechanisms. This approach would require even less effort and fewer resources than running simulations with synthetic subjects. Of course, the feasibility of this simpler idea depends entirely on the model's ability to accurately predict the results of complex social interactions from the outset.

We asked GPT-4 to predict the outcomes of the original baseline experiment from Hoelzemann et al. (2025). The results of this exercise are presented in Appendix Table E3. While direct elicitation occasionally yields reasonable predictions, overall accuracy remains modest. More critically, GPT-4 does not capture the core mechanisms underlying the theory and overlooks the key insight that more data can lead to worse outcomes. By contrast, using AI agents within our experimental framework more effectively replicates the original findings, providing further support for the bottom-up approach. We also find that GPT-4's direct predictions fail to align with the results of several simulation-based extensions, reinforcing the limitations of relying solely on outcome prediction instead of a computational experiment.

6 Discussion

In this paper, we argue that using LLMs as synthetic subjects in simulated experiments offers a powerful new tool for strategy research. We introduce a framework for experimenting with AI agents and apply it to a theory of exploration in the presence of data spillovers. Replicating the online lab experiment from Hoelzemann et al. (2025), we find that LLMs can approximate human behavior in group settings with strategic interdependencies. We then extend the original theory through a series of exploratory simulations. These reveal new and plausible boundary conditions, such as the effects of payoff rivalry and nonstandard agent preferences in reducing herding during search. We also uncover novel potential mechanisms behind the results of Hoelzemann et al. (2025), including social conformity. These insights were neither anticipated nor present in the model's training data, highlighting the value of theorizing with LLMs.

We learned several lessons from this exercise. In our setting, LLMs exhibited a moderate degree of behavioral realism. This is promising, since the key question is whether synthetic subjects' behavior offers a reasonable approximation of how humans would act in the same strategic context.[12] However, the degree of alignment between LLMs and humans is likely context-dependent, as documented by previous studies (Tjuatja et al., 2023). Even in our application, we noted key points of divergence, with LLMs displaying a bias toward rationality. As such, human-subject experiments remain the gold standard for behavioral realism. Furthermore, as others have emphasized in the context of traditional models (Knudsen et al., 2019),

[12]We are indebted to an anonymous reviewer for raising this point and encouraging us to frame AI-agents as synthetic subjects.

LLM-based simulations do not constitute empirical data for hypothesis testing. Rather, their value lies in leveraging the quasi-realism of LLMs to generate theoretically relevant extensions to the starting theory.

Although not directly evident in our application, additional limitations of LLMs warrant closer scrutiny in the context of strategy research. Because LLMs are trained on textual data drawn from the internet, they disproportionately reflect English-speaking, urban populations (Dillion et al., 2023). Unlike the "WEIRD" bias observed in human-subject research (Henrich et al., 2010), this bias is embedded in the foundational model itself, potentially limiting its ability to simulate environments or represent actors with limited digital data. This is especially relevant for subfields such as international business in non-Western settings, or research on privately held companies, including startups and family enterprises. A second concern involves cognitive bias. LLMs exhibit framing and anchoring effects that, while similar to human tendencies, may systematically distort outcomes in strategic experiments (Jones and Steinhardt, 2022).

Despite these issues, a key lesson from our study is that researchers can take concrete steps to ensure the validity of LLM-based simulations. First, borrowing a concept from machine learning, access to prior "hold-out data" was essential for us. It provided a benchmark for evaluating outcomes and a foundation for simulating new experimental conditions. Without this reference point, it would have been difficult to justify extending the analysis to untested scenarios. This approach parallels best practices in agent-based modeling, where researchers typically begin by replicating known results before introducing further complexity (Davis et al., 2007). Second, directly eliciting qualitative explanations from the models proved valuable. These rationales often revealed reasoning patterns similar to those of human subjects, suggesting the alignment was structural rather than coincidental. Still, such responses may reflect hallucination or post-hoc justification. For this reason, researchers' judgment remains critical—not only in extending the starting theory, but also in evaluating the plausibility of model behavior and interpreting the explanations LLMs generate.

We also found that LLM-based simulations offer a moderate degree of control. We were able to introduce meaningful variation to systematically link actions and outcomes, much like in ABMs. Changes to the strategic environment—such as the choice landscape or payoff structure—were easy to implement by direct modification of our script. We also influenced agent behavior by prompting the LLMs toward specific behavioral orientations, such as prosociality. At the same time, the use of a foundation model (in our case, OpenAI's GPT-4) introduces limits to this control. Unlike ABMs, where agents are fully specified by the researcher, LLM agents are partially opaque and not entirely programmable. This complexity may contribute to their behavioral realism, but it also constrains our ability to isolate the mechanisms driving specific outcomes. For this reason, we see an ongoing role for traditional ABMs, particularly when the goal is to identify the essential mechanisms that can explain observed patterns.

Nevertheless, this limited control may offer a distinct advantage for theory development: the potential to surprise the researcher. In ABMs, outcomes reflect only what has been explicitly programmed, so any unanticipated finding still reflects the properties and assumptions specified by the modeler (Knudsen et al., 2019). With LLMs, the internal workings are less transparent, which creates space for truly unexpected behavior to emerge. These surprises can be surfaced by asking LLMs to articulate their reasoning—a new and distinctive capability of this class of models. Even if these rationales are occasionally hallucinatory, they can still be generative, prompting researchers to explore alternative mechanisms or consider extensions to existing theories. We also suspect that some categories of agentic variation may be difficult to represent in traditional ABMs but more readily accessible through LLMs. These include variation in reasoning styles, shifts in moral framing, or sensitivity to social norms. Indeed, most ABMs in our literature review emphasized technical, easily parameterized constructs, such as performance aspirations or risk preferences.

Where might the unique combination of control and realism of AI agents add the most value in strategy research? While the answer is necessarily speculative, our review points to several promising directions. Some domains, such as competitive dynamics and adaptation to environmental change, have long relied on simulation methods. This raises questions about the external validity of their findings. Revisiting these findings with LLM-powered agents that do not rely on prespecified behavioral rules, but instead display elements of human-like reasoning and adaptability, could yield meaningful insights. Other areas, including entrepreneurship and strategic human capital, have yet to fully exploit the variation that simulation enables. Here, LLM-based experiments create new possibilities. For instance, researchers could create AI-managed organizations to evaluate candidate résumés, replicating the design of prior audit studies. They could then manipulate applicant attributes to detect subtle forms of bias that would be costly and time-consuming to identify using human subjects. We also see significant promise in our method being used by researchers running human subject experiments to better design the experiments they run, allowing them to choose manipulations that are more likely to yield real insight. These examples illustrate how the hybrid nature of LLMs—combining computational control with behavioral richness—can open up novel directions for strategy theorizing.

In conclusion, LLM-powered agents may come to play a role similar to that of model organisms in biomedical science. Just as animal models have become a central tool alongside in vitro tests and clinical trials, LLM-based simulations could complement both ABMs and human-subject experiments within strategy. Realizing this vision, however, will require sustained engagement from the strategy research community. In biology, the widespread use of lab mice was made possible by their standardization as a research tool (Hedrich, 2004). Institutions such as The Jackson Laboratory played a critical role by providing widely available mouse strains and establishing best practices for experimental use. In the same spirit, we hope the framework and applications offered here serve as a foundation for standardizing LLM-based simulations, lowering technical barriers and enabling broader adoption by researchers who may not yet be familiar with these tools.

References

Adner, R. and D. A. Levinthal (2024): "Strategy Experiments in Nonexperimental Settings: Challenges of Theory, Inference, and Persuasion in Business Strategy," *Strategy Science*, 9, 311–321.

Agrawal, A., J. S. Gans, and A. Goldfarb (2019): "Exploring the impact of artificial intelligence: Prediction versus judgment," *Information Economics and Policy*, 47, 1–6.

Aher, G., R. I. Arriaga, and A. T. Kalai (2023): "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies," *arXiv preprint*.

Ashokkumar, A., L. Hewitt, I. Ghezae, and R. Willer (2024): "Predicting Results of Social Science Experiments Using Large Language Models," *Working Paper*.

Bail, C. A. (2024): "Can Generative AI improve social science?" *Proceedings of the National Academy of Sciences*, 121.

Bandiera, O., I. Barankay, and I. Rasul (2011): "Field experiments with firms," *Journal of Economic Perspectives*, 25, 63–82.

Banerjee, A. V. (1992): "A simple model of herd behavior," *The Quarterly Journal of Economics*, 107, 797–817.

Billinger, S., N. Stieglitz, and T. R. Schumacher (2014): "Search on rugged landscapes: An experimental study," *Organization Science*, 25, 93–108.

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. (2021): "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*.

Broska, D., M. Howes, and A. van Loon (2024): "The Mixed Subjects Design: Treating Large Language Models as Potentially Informative Observations," *Sociological Methods & Research*, 00491241251326865.

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, et al. (2023): "Sparks of Artificial General Intelligence: Early experiments with GPT-4," *arXiv preprint*.

Carlson, N. and V. Burbano (2025): "The Use of LLMs to Annotate Data in Management Research: Guidelines and Warnings," *Available at SSRN: https://ssrn.com/abstract=4836620*.

Chatterji, A. K., M. Findley, N. M. Jensen, S. Meier, and D. Nielson (2016): "Field experiments in strategy research," *Strategic Management Journal*, 37, 116–132.

Choudhury, P., R. T. Allen, and M. G. Endres (2021): "Machine learning for pattern discovery in management research," *Strategic Management Journal*, 42, 30–57.

Csaszar, F. A., H. Ketkar, and H. Kim (2024): "Artificial Intelligence and Strategic Decision-Making: Evidence from Entrepreneurs and Investors," *Available at SSRN 4913363*.

Dahlander, L. and H. Piezunka (2014): "Open to suggestions: How organizations elicit suggestions through proactive and reactive attention," *Research Policy*, 43, 812–827.

Davis, J. P., K. M. Eisenhardt, and C. B. Bingham (2007): "Developing theory through simulation methods," *Academy of Management Review*, 32, 480–499.

Dawkins, R. (1976): *The Selfish Gene*, Oxford University Press.

Dell'Acqua, F., E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani (2023): "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality," *Harvard Business School TOM Working Paper*.

Di Stefano, G. and C. Gutierrez (2019): "Under a magnifying glass: On the use of experiments in strategy research," *Strategic Organization*, 17, 497–507.

Dillion, D., N. Tandon, Y. Gu, and K. Gray (2023): "Can AI language models replace human participants?" *Trends in Cognitive Sciences*, 27, 597–600.

Doshi, A. R., J. J. Bell, E. Mirzayev, and B. Vanneste (2024): "Generative Artificial Intelligence and Evaluating Strategic Decisions," *Available at SSRN 4714776*.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2024): "GPTs are GPTs: Labor market impact potential of LLMs," *Science*, 384, 1306–1308.

Felin, T. and M. Holweg (2024): "Theory is all you need: AI, human cognition, and causal reasoning," *Strategy Science*.

Felten, E. W., M. Raj, and R. Seamans (2023): "Occupational heterogeneity in exposure to generative ai," *Available at SSRN 4414065*.

Ganco, M. (2017): "NK model as a representation of innovative search," *Research Policy*, 46, 1783–1800.

Gao, Y., D. Lee, G. Burtch, and S. Fazelpour (2025): "Take caution in using LLMs as human surrogates," *Proceedings of the National Academy of Sciences*, 122, e2501660122.

Grimes, M., G. Von Krogh, S. Feuerriegel, F. Rink, and M. Gruber (2023): "From scarcity to abundance: Scholars and scholarship in an age of generative artificial intelligence," *Academy of Management Journal*, 66, 1617–1624.

Grossmann, I., M. Feinberg, D. C. Parker, N. A. Christakis, P. E. Tetlock, and W. A. Cunningham (2023): "AI and the transformation of social science research," *Science*, 380, 1108–1109.

Hagendorff, T. (2024): "Deception abilities emerged in large language models," *Proceedings of the National Academy of Sciences*, 121.

Hagendorff, T., S. Fabi, and M. Kosinski (2023): "Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT," *Nature Computational Science*, 3, 833–838.

Harrison, J. R., Z. Lin, G. R. Carroll, and K. M. Carley (2007): "Simulation modeling in organizational and management research," *Academy of management review*, 32, 1229–1245.

Haveman, H. A. (1993): "Follow the leader: Mimetic isomorphism and entry into new markets," *Administrative science quarterly*, 593–627.

Haynes, W. A., A. Tomczak, and P. Khatri (2018): "Gene Annotation Bias Impedes Biomedical Research." *Scientific Reports*, 8.

Hedrich, H. (2004): *The laboratory mouse*, Academic Press.

Henrich, J., S. J. Heine, and A. Norenzayan (2010): "The weirdest people in the world?" *Behavioral and Brain Sciences*, 33, 61–83.

Hoelzemann, J., G. Manso, A. Nagaraj, and M. Tranchero (2025): "The streetlight effect in data-driven exploration," *NBER Working Paper w32401*.

Horton, J. and R. Horton (2024): "EDSL: Expected Parrot Domain Specific Language for AI Powered Social Science," Whitepaper, Expected Parrot.

Horton, J. J. (2023): "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" *NBER Working Paper w31122*.

Jia, N., X. Luo, Z. Fang, and C. Liao (2024): "When and how artificial intelligence augments employee creativity," *Academy of Management Journal*, 67, 5–32.

Jones, E. and J. Steinhardt (2022): "Capturing failures of large language models via human cognitive biases," *Advances in Neural Information Processing Systems*, 35, 11785–11799.

Kauffman, S. A. (1993): *The Origins of Order: Self-organization and Selection in Evolution*, Oxford University Press.

Knudsen, T. (2024): "Agent-Based Modelling for Strategy," in *The Oxford Handbook of Agent-based Computational Management Science*, Oxford University Press.

Knudsen, T., D. A. Levinthal, and P. Puranam (2019): "A model is a model," *Strategy Science*, 4, 1–3.

Krieger, J. L. (2021): "Trials and terminations: Learning from competitors' R&D failures," *Management Science*, 67, 5525–5548.

LaFollette, H. and N. Shanks (1995): “Two models of models in biomedical research,” *The Philosophical Quarterly (1950-)*, 45, 141–160.

Lampinen, A. K., I. Dasgupta, S. C. Chan, H. R. Sheahan, A. Creswell, D. Kumaran, J. L. McClelland, and F. Hill (2024): “Language models, like humans, show content effects on reasoning tasks,” *PNAS nexus*, 3.

Lazer, D. and A. Friedman (2007): “The network structure of exploration and exploitation,” *Administrative science quarterly*, 52, 667–694.

Leiblein, M. J., J. J. Reuer, and T. Zenger (2018): “What makes a decision strategic?” *Strategy Science*, 3, 558–573.

Lerner, J. and R. Nanda (2020): “Venture capital’s role in financing innovation: What we know and how much we still need to learn,” *Journal of Economic Perspectives*, 34, 237–261.

Levine, S. S., O. Schilke, O. Kacperczyk, and L. G. Zucker (2023): “Primer for experimental methods in organization theory,” *Organization Science*, 34, 1997–2025.

Levinthal, D. A. (1997): “Adaptation on rugged landscapes,” *Management Science*, 43, 934–950.

Lindsey, J., W. Gurnee, E. Ameisen, B. Chen, A. Pearce, N. L. Turner, et al. (2025): “On the Biology of a Large Language Model,” *Transformer Circuits Thread*.

Luo, X., A. Rechardt, G. Sun, K. K. Nejad, F. Yáñez, et al. (2025): “Large language models surpass human experts in predicting neuroscience results,” *Nature human behaviour*, 9, 305–315.

Makadok, R., R. Burton, and J. Barney (2018): “A practical guide for making theory contributions in strategic management,” *Strategic Management Journal*, 39, 1530–1545.

Manning, B. S., K. Zhu, and J. J. Horton (2024): “Automated social science: Language models as scientist and subjects,” *NBER Working Paper w32381*.

March, J. G. (1991): “Exploration and exploitation in organizational learning,” *Organization science*, 2, 71–87.

Mei, Q., Y. Xie, W. Yuan, and M. O. Jackson (2024): “A Turing test of whether AI chatbots are behaviorally similar to humans,” *Proceedings of the National Academy of Sciences*, 121.

Mueller, J. (2018): “Finding new kinds of needles in haystacks: Experimentation in the course of abduction,” *Academy of Management Discoveries*, 4, 103–108.

Ott, T. E. and D. P. Hannah (2024): “On the Origin of Entrepreneurial Theories: How Entrepreneurs Craft Complex Causal Models with Theorizing and Data,” *Strategy Science*.

Park, J. S., J. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein (2023): “Generative Agents: Interactive Simulacra of Human Behavior,” *arXiv preprint arXiv:2304.03442*.

Puranam, P. and M. Swamy (2016): “How initial representations shape coupled learning processes,” *Organization Science*, 27, 323–335.

Rumelt, R. P., D. Schendel, and D. J. Teece (1991): “Strategic management and economics,” *Strategic Management Journal*, 12, 5–29.

Sarstedt, M., S. J. Adler, L. Rau, and B. Schmitt (2024): “Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines,” *Psychology & Marketing*, 41, 1254–1270.

Shrestha, Y. R., V. F. He, P. Puranam, and G. von Krogh (2021): “Algorithm supported induction for building theory: How can we use prediction models to theorize?” *Organization Science*, 32, 856–880.

Si, C., D. Yang, and T. Hashimoto (2024): “Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers,” *arXiv preprint*.

Siggelkow, N. (2011): “Firms as systems of interdependent choices,” *Journal of Management Studies*, 48, 1126–1140.

Srivastava, A., A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, et al. (2022): “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models,” *arXiv preprint arXiv:2206.04615*.

Tjuatja, L., V. Chen, S. T. Wu, A. Talwalkar, and G. Neubig (2023): “Do LLMs exhibit human-like response biases? A case study in survey design,” *arXiv preprint*.

Tranchero, M. (2025): “Data-driven search and innovation: Evidence from genome-wide association studies,” *University of Pennsylvania*.

Tu, X., J. Zou, W. Su, and L. Zhang (2024): "What Should Data Science Education Do With Large Language Models?" *Harvard Data Science Review*, 6.

Van den Steen, E. (2018): "The strategy in competitive interactions," *Strategy Science*, 3, 574–591.

Vuculescu, O. (2017): "Searching far away from the lamp-post: An agent-based model," *Strategic Organization*, 15, 242–263.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. (2022): "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, 35, 24824–24837.

A Review of Computational and Human Subject Experiments in SMJ

In this Appendix, we summarize the methodology used to identify experimental studies published in the *Strategic Management Journal* (SMJ), including both agent-based models (ABMs) and human subjects experiments. We then describe how these studies were classified by research topic and source of variation. While not intended to be exhaustive, this review seeks to provide a systematic overview of the questions that strategy scholars explore and the methodological tools they currently employ to identify causal mechanisms.

A.1 Compiling List of Studies

We obtained the list of studies from Scopus, an academic database curated by Elsevier. Specifically, we used Scopus' advanced query functionality to search titles, abstracts, and author-supplied keywords for all articles published in SMJ. Our first query targeted experimental studies using human subjects, including both field and laboratory experiments:

TITLE-ABS-KEY("field experiment") OR TITLE-ABS-KEY("RCT") OR TITLE-ABS-KEY("randomized controlled trial") OR TITLE-ABS-KEY("randomised controlled trial") OR TITLE-ABS-KEY("lab") OR TITLE-ABS-KEY("laboratory") OR TITLE-ABS-KEY("controlled experiment") AND SRCTITLE("Strategic Management Journal") AND LIMIT-TO(EXACTSRCTITLE, "Strategic Management Journal")

This search returned 44 articles. After manually filtering out papers that were not actually experimental (e.g., special issue introductions or methodological pieces), we retained 32 studies for analysis. Our second query aimed to identify studies using computational experiments, including agent-based models, system dynamics, and other simulation-based methods:

TITLE-ABS-KEY("agent based model") OR TITLE-ABS-KEY("computer simulation") OR TITLE-ABS-KEY("simulated experiment") OR TITLE-ABS-KEY("simulation") OR TITLE-ABS-KEY("computational simulation") OR TITLE-ABS-KEY("computational model") OR TITLE-ABS-KEY("nk model") OR TITLE-ABS-KEY("nk fitness") OR TITLE-ABS-KEY("system dynamics") OR TITLE-ABS-KEY("genetic algorithms") OR TITLE-ABS-KEY("cellular automata") OR TITLE-ABS-KEY("nk landscape") OR TITLE-ABS-KEY("fitness landscape") OR TITLE-ABS-KEY("abm") AND SRCTITLE("Strategic Management Journal") AND LIMIT-TO(EXACTSRCTITLE, "Strategic Management Journal")

This search returned 83 articles. After manually excluding studies that did not use computational modeling techniques, we retained 49 studies. The final list of studies is shown below in Appendix Table A1.

A.2 Classifying Studies by Topic

Next, after reviewing the studies listed above, we inductively derived eight distinct topic categories based on clusters of related papers. While we acknowledge that this is inherently a subjective exercise and that some overlap between categories is inevitable, it nonetheless highlights meaningful groupings of research activity within the literature. In total, we derived eight distinct topic categories:

1. Competitive Strategy: How firms compete with each other, and why some perform better than others.

2. Corporate Strategy: How firms create value within the organization, including decisions about resource allocation, diversification, and strategy implementation.

3. Adaptation & Environmental Change: How firms adjust their strategies and positioning in response to changes in the external environment, be it technological, competitive, or due to other shifts.

4. Cognition & Decision-Making: How behavioral microfoundations, such as perception, cognition, and emotion, influence strategic decision-making.

Table A1: List of SMJ Studies Included in the Systematic Review

Simulations	Experiments
Srikanth and Ungureanu, 2025	Novelli and Spina, 2024
Sharapov and Ross, 2023	Santamaria et al., 2024
Ketkar and Workiewicz, 2022	Hurst et al., 2024
Martignoni and Keil, 2021	Boulongne et al., 2024
Dong, 2021	Healey and Hodgkinson, 2024
Davis and Aggarwal, 2020	Wang et al., 2023
Posen et al., 2020	Kotha et al., 2023
Chen et al., 2019	Dahlander et al., 2023
Posen and Martignoni, 2018	Richter et al., 2023
Kim and Anand, 2018	Mickeler et al., 2023
Posen et al., 2018	Durand and Huysentruyt, 2022
Luoma et al., 2017	Nai et al., 2022
Aggarwal et al., 2017	Lane et al., 2021
Csaszar and Levinthal, 2016	Tong et al., 2021
Lee and Puranam, 2016	Billinger et al., 2021
Rahmandad and Repenning, 2016	Hasan and Koning, 2019
Lee et al., 2016	Chatterji et al., 2019
Martignoni et al., 2016	Flammer and Kacperczyk, 2019
Andersen and Bettis, 2015	Meier et al., 2019
Keyhani et al., 2015	Laureiro-Martínez and Brusoni, 2018
Miller and Lin, 2015	Christensen et al., 2017
Sakhartov and Folta, 2015	Elfenbein et al., 2017
Baumann and Stieglitz, 2014	Chatterji et al., 2016
Fang et al., 2014	Døjbak et al., 2016
Baum et al., 2014	Phadnis et al., 2015
Kogut et al., 2014	Jones et al., 2015
Schilling and Fang, 2014	Cain et al., 2015
Knudsen et al., 2014	Di et al., 2014
Garcia-Sanchez et al., 2014	Robert et al., 2011
Posen et al., 2013	Knez and Camerer, 1994
Markle, 2011	Brockner et al., 1993
Hu et al., 2011	
Aggarwal et al., 2011	
Miller et al., 2009	
Andersen et al., 2007	
Gavetti et al., 2005	
Gary, 2005	
Miller and Arikan, 2004	
Knott, 2003	
Zott, 2003	
Johnson and Hoopes, 2003	
Crossland and Smith, 2002	
Lee et al., 2002	
Adner, 2002	
Levy, 1994	
Mezias and Glynn, 1993	
Istvan, 1992	
Merten, 1991	
Morecroft, 1984	

5. Strategic Human Capital: How firms attract, develop, and manage talent to create value.

6. Entrepreneurship: How new ventures are created, and how individuals generate value, innovate, and disrupt industries.

7. Organizational Knowledge: How knowledge is accumulated, applied, and transferred within and between organizations.

8. Networks: How the structure of connections among individuals, organizations, or institutions shapes strategy.

We then manually classified each of the studies in Appendix Table A1 into one of the eight topic categories based on the author-supplied keywords and titles, and in ambiguous cases the abstracts or introductions. To demonstrate how these classifications were applied, we provide illustrative examples in Appendix Table A2.

Table A2: Example Studies by Topic Category

Topic	Example
Competitive Strategy	*Time delays, competitive interdependence, and firm performance* (Luoma et al., 2017)
Corporate Strategy	*Implementation strategy and performance outcomes in related diversification* (Gary, 2005)
Adaptation & Environmental Change	*Adaptive capacity to technological change: A microfoundational approach* (Aggarwal et al., 2017)
Cognition & Decision Making	*Mental representation and the discovery of new strategies* (Csaszar and Levinthal, 2016)
Strategic Human Capital	*The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance* (Tong et al., 2021)
Entrepreneurship	*Demand pull versus resource push training approaches to entrepreneurship: A field experiment* (Santamaria et al., 2024)
Organizational Knowledge	*Engineering serendipity: When does knowledge sharing lead to knowledge production?* (Lane et al., 2021)
Networks	*Does evidence of network effects on firm performance in pooled cross-section support prescriptions for network strategy?* (Baum et al., 2014)

A.3 A Taxonomy of Experimental Sources of Variation

Finally, we review the full set of studies to identify which sources of variation each one leverages. In simulation papers, these sources often take the form of alternative initial conditions, model parameters, or exogenous shocks, whereas in experimental studies, they often correspond to the randomized interventions introduced by researchers. Observing recurring themes across papers, we inductively developed a set of categories to classify these sources of variation. Note that not every study will experimentally manipulate all of them.

The first three categories relate to the *agents* studied, such as individual decision-makers or organizations.

1. Agent Objectives: What agents are trying to achieve. This includes preferences (e.g., utility functions or risk tolerance), specific performance aspirations, and the nature of their goals (e.g., maximizing profit or promoting social welfare).

2. Agent Capabilities: What agents have at their disposal to achieve their objectives. This includes any resource endowments, internal abilities (e.g., the capacity to explore a larger search space), and private knowledge about the environment.

3. Agent Decision Rules: How agents pursue their objectives leveraging their capabilities. This includes decision-making rules, heuristics, routines, algorithms (e.g., greedy search or softmax), and other behavioral policies.

The remaining six categories relate to the environments in which agents operate.

1. Population: The number of agents active in the environment.

2. Time Horizon: The temporal structure of the environment, such as the number of periods or rounds.

3. Choice Landscape: The nature of the choices available to agents. For example, in NK models, this corresponds to N (the number of choice variables) and K (the degree of interdependence among them).

4. Payoff Structure: The rewards associated with each choice. This can simply be the magnitudes of earnings, or can include complex features like probabilistic payoffs or payoff interdependence across agents' choices.

5. Information Set: The degree to which relevant aspects of the environment are public knowledge for the agents.

6. Spatial Structure: The topological arrangement of the landscape (e.g., the correlation in payoff across similar choices), which could be geographical or network-based.

For each study, we record whether there is any variation employed for each given category. We provide an illustrative example for one such study in the Appendix Table A3 below.

Table A3: Coding the Sources of Variation for Csaszar and Levinthal (2016)

Dimension	**Coded Value**	**Explanation**
Agent Objectives	0	Managers' only objective is to maximize performance metrics
Agent Capabilities	1	They vary the strength of managers' cognitive limit (M')
Agent Decision Rules	1	They vary the search strategy (e.g., exploration parameter ε_p)
Population	0	The population is held fixed
Time Horizon	1	They show how search changes over 2000 periods
Choice Landscape	1	They vary the interdependence of choices (K)
Payoff Structure	1	They vary the contribution of rewards to profits (w)
Information Set	0	The information set is held fixed
Spatial Structure	0	There are no topological features

B Current Applications of LLMs in Research

Large Language Models (LLMs) have captured the attention of the public because of their impressive performance in a variety of domains (Bubeck et al., 2023), such as image generation (Qu et al., 2023) and computer programming (Kazemitabaar et al., 2024). Increasingly, LLMs are being used as tools for academic research, with their versatility proving to be useful for various tasks (Grimes et al., 2023). We list the main use cases of LLMs in research in the Appendix Table B1.

Table B1: Outline of How LLMs are Being Used in Academic Research

Application	**Description**	**Example(s)**
Dataset Cleaning	Preprocess and clean datasets	Chong et al. (2022)
Data Analysis	Run statistical methods, including regressions	Ma et al. (2023)
Data Annotation	Annotate textual data and create labels mapping into theoretical constructs	Carlson and Burbano (2025)
Data Visualization	Create visual representations of data	Ye et al. (2024)
Scientific Writing	Refine language, including drafting full abstracts and manuscripts	Salvagno et al. (2023)
Literature Reviews	Automate search for articles	Agarwal et al. (2024)
Summarization	Generates accessible summaries of scientific content	Anderson et al. (2023)
Refining Research Instruments	Design and refining tools like survey questionnaires	Grimes et al. (2023); Charness et al. (2023)
Survey Respondents	Use LLMs as survey respondents	Argyle et al. (2023); Brand et al. (2023); Li et al. (2024)
Replicate Experiments	Replicate known behavioral experiments, like the Milgram shock experiment	Aher et al. (2023); Ashokkumar et al. (2024); Horton (2023)
Hypothesis Generation	Ask LLMs to formulate new testable hypotheses or research ideas	Doshi and Hauser (2024); Si et al. (2024)

Several uses of LLMs leverage their advanced capabilities with data management. Researchers have leveraged LLMs to clean datasets (Chong et al., 2022), perform data analysis (Ma et al., 2023), annotate textual data (Carlson and Burbano, 2025), and generate visualizations of key variables and their relationships (Ye et al., 2024). Beyond data tasks, LLMs have been used extensively in scientific writing, even for generating abstracts and drafts. For instance, it is estimated that up to 20% of the content in computer science conference papers is now substantially AI-modified (Liang et al., 2024). Among other use cases, LLMs have also been leveraged to conduct systematic literature reviews (Agarwal et al., 2024) and refine research instruments such as survey questionnaires (Grimes et al., 2023; Charness et al., 2023).

Besides these applications, there is one feature of LLMs that may prove to be particularly useful for academic research: they have a remarkable propensity to exhibit human-like traits. For example, a recent study by Mei et al. (2024) subjected LLM chatbots to a personality test and a series of behavioral games, finding that their responses were statistically indistinguishable from randomly picked human subjects. Another body of research finds that LLMs rely on human-like heuristics and are prone to similar cognitive errors, reasoning, and even moral judgment (Dillion et al., 2023; Lampinen et al., 2024; Hagendorff, 2024). Given these similarities, a growing number of researchers are seeking to understand whether there are insights we can learn about humans just by indirectly studying the behavior of LLMs (Bail, 2024; Grossmann et al., 2023). In this context, LLMs would not be inputs into the research process, but rather the subjects of study, serving as proxies for humans (Manning et al., 2024; Park et al., 2023). While LLMs do not perfectly mirror human decision-making (Tjuatja et al., 2023; Mohammadi, 2024), they might provide a flawed but potentially very useful approximation as “homo silicus” (Horton, 2023).

Indeed, an active body of literature is developing around this promise. Scholars have tried to use LLMs as survey respondents to uncover consumer demand (Brand et al., 2023; Li et al., 2024) and predict voting behavior (Argyle et al., 2023). However, most research to date has focused on replicating lab-based experiments using LLMs in the place of human subjects. Aher et al. (2023) and Ashokkumar et al. (2024) replicate a suite of behavioral experiments, from the wisdom of crowds to the Milgram shock experiment, finding results consistent with past studies using human subjects. Horton (2023) investigates the strategic capabilities of LLMs by having them participate in games like the prisoner's dilemma or the dictator game. While these games involve two players only, more recent studies have incorporated more players and richer design elements, like information deficiencies and spatial reasoning (Wu et al., 2023; Xu et al., 2023).

Finally, some studies are using LLMs for research ideation and direct hypothesis generation. Doshi and Hauser (2024) find that when fiction writers use generative AI to obtain ideas for a story, the narratives are evaluated as more novel but tend to be more similar to each other. Similar results are found by Si et al. (2024) when tasking an LLM with generating new research ideas and comparing them with human experts. Manning et al. (2024) study automated hypothesis generation employing LLMs to propose potential causal relationships and design experiments to test those relationships.

C Additional Information on EDSL

C.1 What is EDSL?

Expected Parrot Domain-Specific Language (EDSL) is an open-source Python library designed to streamline and automate AI-powered research workflows, particularly those requiring large-scale coordination of LLMs (Horton and Horton, 2024). At its core, EDSL enables structured interactions with the language models: each model receives a scenario-specific, parameterized question, and their response is returned as a formatted result. We take advantage of this structure to replicate the laboratory experiment in Hoelzemann et al. (2025). Since all interactions relied on APIs, EDSL supplied the necessary guardrails to format inputs and parse outputs reliably. Beyond our use case, EDSL is also tailored for applications such as:

- Designing and administering surveys and experiments
- Performing data labeling and content moderation
- Conducting computational social science and market research
- Gathering and analyzing responses from both humans and AI systems

C.2 Example EDSL script

We used multiple variations of scripts to simulate different configurations of the experiment, depending on the specific assumptions employed (e.g., payoff ambiguity and rivalry). Below, we provide an example for the specific case of non-rivalry and non-ambiguity. Note that many aspects of the environment are parameters manipulated by the researcher, whose precise value will depend on the specific simulation.

Instructions

General Information:
Welcome. This is an experiment in the economics of decision-making. If you pay close attention to these instructions, you can earn a significant amount of money paid to you at the end of the experiment. Following these instructions, you will be asked to make some choices. There are no correct choices. Your choices depend on your preferences and beliefs, so different participants will usually make different choices. You will be paid according to your choices, so read these instructions carefully and think before you decide.

The Basic Idea:
There are `{len(MOUNTAINS_DISTRIBUTION)}` mountains and each of them hides one type of gem, which can only be found by exploring the mountain. There are 3 types of gems hidden in the `{len(MOUNTAINS_DISTRIBUTION)}` mountains: Diamonds, Rubies, and Topazes. The exact values of the topazes, rubies, and diamonds vary across rounds but the diamonds are always worth more than the rubies and the rubies are always worth more than the topazes. You choose which mountains to explore and the value of the gems you find are your earnings in dollars. Your objective is to maximize your own earnings.

Location of Gems:
`{location_of_gems}`

The Mapped Mountain:
At the beginning of each round, one mountain will be randomly selected to be mapped and its gem value will be revealed to all participants. Each participant will be able to see the same gem contained by the mountain. The mountain chosen for mapping is random and changes in each round. Besides the value of the mapped mountain, no participant has any other initial information in Stage 1 on the

location of gems.

How Participants Choose Mountains:
There are `{no_of_players}` participants including you in total. In each round, participants choose which mountain to explore. The choice does not happen simultaneously, but participants choose sequentially, one after the other, according to a random order. You can choose to explore any mountain you wish or select the mapped mountain. If you choose the same mountain chosen by other participants, each of you will receive the full value uncovered. Similarly, if someone else chooses the same mountain that you previously chose, you will still receive the full gem's value (and so will the other participant(s) who chose it). This means that payoffs are non-rival and there is no penalty in choosing the same mountain as other players. To repeat, no participant has any private information in Stage 1 on the location of the gems, besides the common knowledge about the mapped mountain.

Each Round Has 2 Stages:
A round consists of 2 stages. At the beginning of a new round, gems are randomly allocated to the `{len(MOUNTAINS_DISTRIBUTION)}` different mountains. The position of gems will not be reset between the two stages in a round. Then, before Stage 1 begins, one mountain will be mapped and its value revealed to everyone. In Stage 1, all participants sequentially choose one mountain to explore. Before choosing a mountain, you will see which mountains have been selected by the other participants in your group who chose before you. You can choose the same mountain chosen by other participants or a different mountain. At the end of Stage 1, the gems hidden in each mountain selected by all participants in Stage 1 are revealed, and you earn the value of the gem hidden in the mountain you chose. In Stage 2, you can again choose any of the same `{len(MOUNTAINS_DISTRIBUTION)}` mountains; that is, you can either choose the same mountain from Stage 1 or switch to another one. The position of gems remains the same as in Stage 1, but this time you will also see the gems located in the mountains revealed in Stage 1 in addition to the mapped mountain. At the end of Stage 2, the gems hidden in each mountain selected by all participants in Stage 2 are revealed, and you earn the value of the gem hidden in the mountain you chose in Stage 2. Your total earnings for the round equal the sum of the value of the gem you found in Stage 1 and the value of the gem you found in Stage 2. Again, if multiple players choose the same mountain, they all receive its full value.

Payment:
At the end of the round, you will be paid an amount equivalent to the sum of payoffs you earned in Stage 1 and Stage 2. This protocol of determining payments suggests that you should choose in each Stage knowing that your choice directly determines your payment because the dollar value of the gems you select will directly translate into your earnings.

Frequently Asked Questions:

- Q1: Is this some kind of psychology experiment with an agenda you haven't told us?
  A: No, it is an economics experiment. These instructions are meant to clarify how you earn money and our interest is in seeing how people make decisions.
- Q2: Is there a "correct" or "wrong" choice of action? Is this kind of a test?
  A: No, your optimal choice depends on your preferences and beliefs and different people may hold different beliefs.
- Q3: Will there be any mountains with empty payoffs (no gem at all)?
  A: No, each and every mountain contains a gem and you are guaranteed a payoff for choosing any mountain whether it is a topaz, ruby, or a diamond.

- Q4: Will there be any negative payoffs for some hidden mountains?
  A: No, there is no potentially lower payoff than of a topaz, which is always positive.
- Q5: Are the payoffs split if more players choose the same mountain?
  A: No, each participant receives the full value of the gem uncovered, regardless of how many participants choose a mountain in that specific stage.

Instructions for the Risk Preference:
The Choice: You will be asked to choose between two options, "Option A" and "Option B" where:
"Option A" always pays $4.00 with probability p and $3.20 otherwise.
"Option B" always pays $7.70 with probability p and $0.20 otherwise.

Repeated Choices:
You will be asked to make a choice between "Option A" and "Option B" not once, but ten times where p will increase from 10% to 100%, 10% at a time.
For example, the first choice will have p=10% and you will choose whether you prefer "Option A" ($4.00 with a 10% chance or $3.20 otherwise) or "Option B" ($7.70 with a 10% chance or $0.20 otherwise).
Each successive choice will increase p by 10 percentage points until the last choice where "Option A" will pay $4.00 with certainty, and "Option B" will pay $7.70 with certainty.
Note: Once you switch from choosing "Option A" to "Option B", it makes sense that you will continue to choose "Option B" in all consecutive choice problems. For example, if you prefer "Option B" when p=80%, then it makes sense to prefer "Option B" when p=90% and when p=100%, since "Option B" is even more attractive in these choice problems.

Payment for risk preference task:
The computer will randomly select one of the 10 choice problems and pay you according to your choice in that problem where the computer will decide the outcome based on the value of p.

C.3 Link to Code Repository

In the interest of transparency and open science, we provide the full code used to simulate the experiments. The implementation is available in a public GitHub repository. We include two Jupyter notebooks corresponding to different information conditions in the game:

- **No-data condition:** Participants make decisions without observing the value of any alternative.
  https://github.com/arulmabr/theorizing_with_llms/blob/main/notebooks/streetlight_no-data.ipynb
- **Data condition:** Participants observe the value of one alternative at the outset, either low, medium, or high in value.
  https://github.com/arulmabr/theorizing_with_llms/blob/main/notebooks/streetlight_with-data.ipynb

D Key Parameters for LLM-Powered Experiments

Environment: The researcher must specify the common set of conditions that the agent(s) encounters during an in silico experiment. The following environmental dimensions are especially relevant to strategic management:

1 **Population:** The number of agents active in the environment. The computational environment theoretically supports an unlimited number of agents.

2 **Time Horizon:** The temporal structure of the environment, such as the number of periods or rounds. The computational environment theoretically supports an unlimited number of rounds.

3 **Choice Landscape:** The nature of the choices available to agents. The computational environment theoretically supports an unlimited number of choices, or combinations of choices, enabling the tunable interaction complexity seen in NK models (Kauffman, 1993; Levinthal, 1997).

4 **Payoff Structure:** The rewards associated with each choice. This can simply be the magnitudes of earnings, or can include complex features like probabilistic payoffs or payoff rivalry across agents. The computational environment theoretically supports unlimited combinations of earning schedules.

5 **Information Set:** The degree to which relevant aspects of the environment are public knowledge. The computational environment make this straightforwardly tunable, with can be done in the initial instructions. For instance, agents may be told the options available, but not the payoffs associated with each option.

6 **Spatial Features:** The topological arrangement of the landscape, which could be geographical or network-based. The computational environment theoretically supports unlimited spatial arrangements of agents and choices.

Agents: The researcher must also specify the characteristics of each agent. The following agentic dimensions are especially relevant to strategic management:

1 **Objectives:** What agents are trying to achieve. This includes preferences (e.g., risk tolerance), specific performance aspirations, and the nature of their goals - such as maximizing profit or promoting social welfare. LLMs can be tuned to pursue different objectives by prompting them with additional scripts. They can also be assigned specific roles, which might contain implicit objectives that the language model needs to infer. For example, a "student LLM" might focus on maximizing learning, while a "firm LLM" might choose to maximize profits.

2 **Capabilities:** What agents have at their disposal. This includes any resource endowments, internal abilities (e.g., the capacity to explore a larger search space), and private knowledge about the environment. Once the agent is informed of these capabilities, they must be encoded throughout the implementation to ensure the engine accurately tracks and constrains the agent's behavior.

3 **Decision-Rules:** How agents pursue their goals. This includes decision-making rules, heuristics, algorithms (e.g., greedy search, softmax), and other behavioral policies. This once more entails prompting them with additional scripts.

Turn-taking Procedures: The researcher must specify how decisions are made by agents. There are two primary methodologies they can choose from:

1 **Sequential Decision-Making:** Agents are allotted a "turn" to make their decision. After each turn, the state space is updated and the experiment proceeds until a predetermined endpoint has been reached. The researcher would need to decide how turns are allotted, e.g., using a random order or perhaps having an independent LLM interact with each agent, and then decide (Manning et al., 2024).

2 **Simultaneous Decision-Making:** Agents take actions at the same time within a round. After each round, the state space is updated and the experiment proceeds until a predetermined endpoint has been reached. The researcher would simply need to decide the number of rounds (Huang et al., 2024).

Communication Channels: The researcher must decide whether the AI agents can speak with each other during the course of the experiment (or engage in other forms of interaction, such as trading). This functionality must be programmed into the engine but remains feasible with LLMs (Xu et al., 2023).

Outcomes: The researcher must specify which outcomes to record as a function of the research question and the theoretical constructs explored.

1 **Quantitative Data:** This includes participant choices, their earnings as a function of collective decisions, the time or effort taken to make a decision, or any other measurable outcome. The engine is also able to calculate more complex corollary measures. This data could then be analyzed using traditional econometric techniques for rigorous inference.

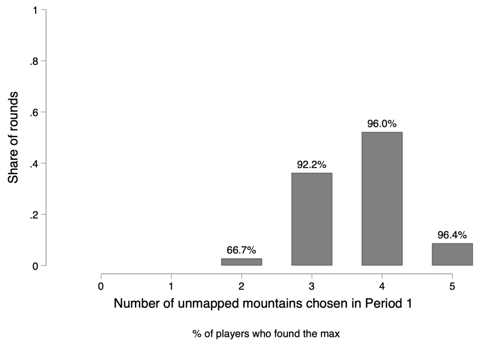
2 **Qualitative Responses:** This primarily includes the explanations that AI agents provide for their decisions, but could include anything in light of LLMs' language comprehension and generation capabilities (e.g. their subjective experiences navigating a certain environment). Collecting these responses is essential for eliciting mechanisms (see section 2.2), as well as for verifying the experiments are working as intended (e.g. seeing whether an agent is embodying an assigned role).

Interventions: These are isolated changes to any element of the game environment or the agents themselves. With LLMs, it becomes possible to pull levers that are not normally available in real-world lab experiments, such as manipulating an agent's preferences or tolerance for risk. Once all the building blocks have been specified, the researcher can begin simulating the experiment, adjusting one element at a time and tracing how the outcomes of interest evolve.

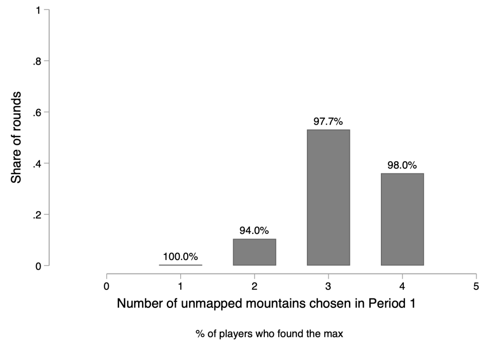
E Additional Figures and Tables

Figure E1: Number of Unknown Mountains Chosen in Period 1 by Human Subjects

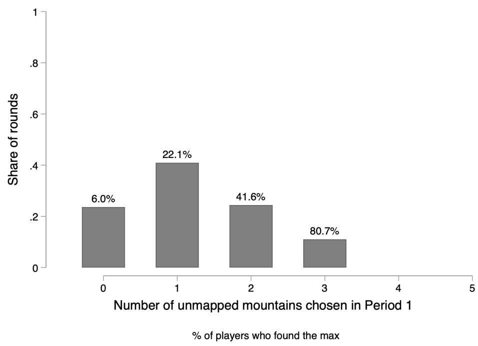
(i) No-data condition



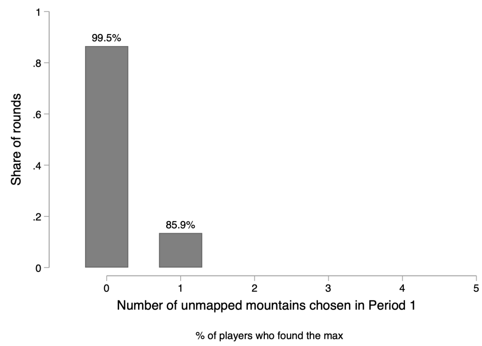
(ii) Low-value condition



(iii) Medium-value condition

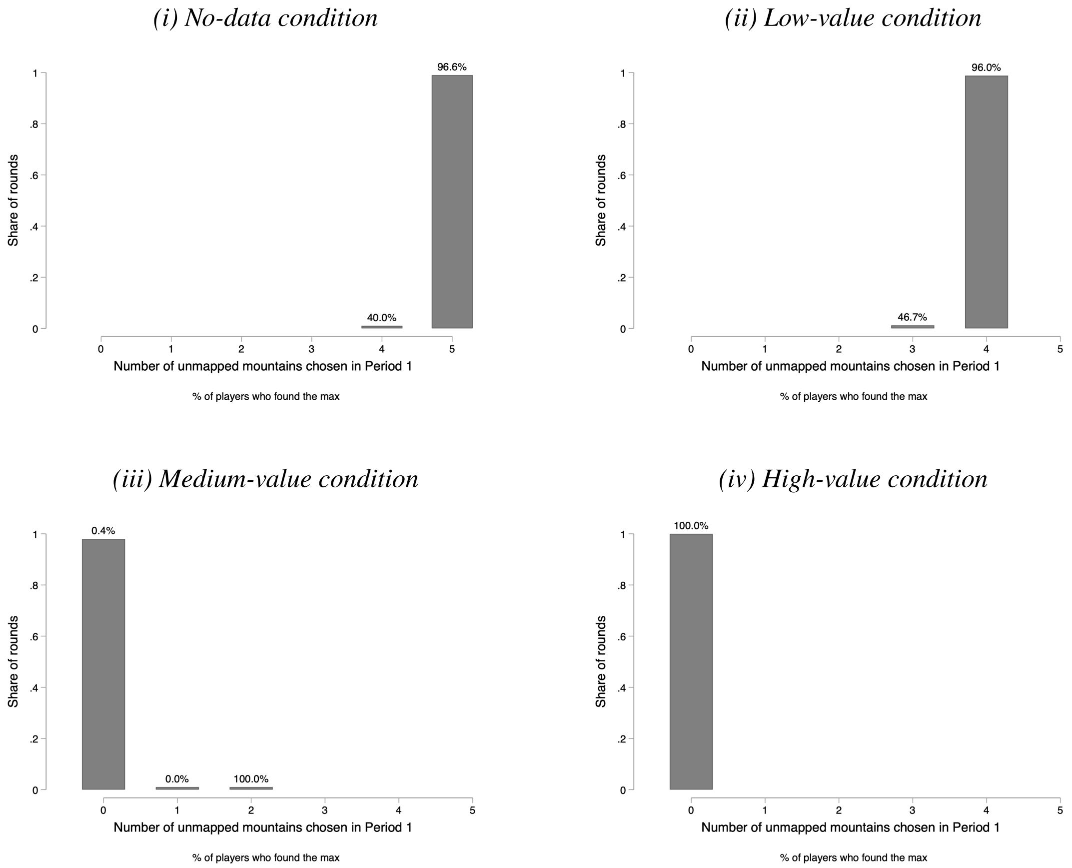


(iv) High-value condition



Note: Each plot represents the empirical frequencies of rounds for each possible number of unknown options chosen in period 1, shown separately by experimental condition. The text written on top of each bar shows the share of participants who found the maximum payoff in period 2. The figure is reproduced with permission from Hoelzemann et al. (2025).

Figure E2: Number of Unknown Mountains Chosen in Period 1 by AI Agents (Baseline)



Note: Each plot represents the empirical frequencies of rounds for each possible number of unknown options chosen in period 1, shown separately by experimental condition. The text written on top of each bar shows the share of participants who found the maximum payoff in period 2.

Table E1: Quantitative Results of Experiment with Human Subjects

Panel A: Round-level Outcomes

	Individual payoff	I(Individual found max)	I(Group found max)
	(1)	(2)	(3)
High	6.682***	0.039**	0.003
	(0.143)	(0.011)	(0.006)
Low	0.889***	0.037**	0.006
	(0.096)	(0.011)	(0.007)
Medium	-2.261***	-0.645***	-0.539***
	(0.214)	(0.028)	(0.039)
Constant	13.349***	0.968***	1.012***
	(0.163)	(0.018)	(0.020)
Round order FE	Yes	Yes	No
Block order FE	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes
Observations	7000	7000	1400

Panel B: Analysis of Mechanisms

	Exploration	Individual payoff		I(Individual found max)		I(Group found max)	
	Round	Period 1	Period 2	Period 1	Period 2	Period 1	Period 2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
High	-75.059***	6.499***	0.191*	0.784***	0.037**	0.310***	0.003
	(1.770)	(0.097)	(0.073)	(0.008)	(0.011)	(0.015)	(0.006)
Low	5.744***	0.664***	0.225**	0.055***	0.036**	0.122***	0.006
	(1.300)	(0.074)	(0.062)	(0.007)	(0.011)	(0.020)	(0.007)
Medium	-34.130***	1.251***	-3.511***	-0.134***	-0.644***	-0.444***	-0.539***
	(2.450)	(0.122)	(0.142)	(0.007)	(0.027)	(0.029)	(0.039)
Constant	83.977***	3.610***	9.717***	0.187***	0.963***	0.752***	1.012***
	(2.045)	(0.104)	(0.117)	(0.008)	(0.018)	(0.018)	(0.020)
Round Order FE	No	Yes	Yes	Yes	Yes	No	No
Block order FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1400	7000	7000	7000	7000	1400	1400

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. Estimates from OLS models. In Panel A, the sample in Columns 1 and 2 is at the participant-round level (5 participants $\times$ 1400 rounds). The sample in Column 3 is at the group-round level (1400 rounds). *Individual payoff*= participant-level round payoffs in Canadian dollars; *I(Individual found max)*:0/1=1 if the location of the maximum was found by the participant; *I(Group found max)*:0/1=1 if the location of the maximum was found by at least one participant in the round. The excluded category captured by the constant is the condition without data.

In Panel B, the sample in Column 1 is at the group-round level (1400 rounds). The sample in Columns 2, 3, 4, 5 is at the participant-period level (5 participants $\times$ 1400 periods of each type). The sample in Columns 6 and 7 is at the group-period level (1400 periods of each type). *Exploration*= share of unknown mountains explored in the round; *Individual payoff*= participant-level period payoffs in Canadian dollars; *I(Individual found max)*:0/1=1 if the location of the maximum was found by the participant in the period; *I(Group found max)*:0/1=1 if the location of the maximum was found by at least one participant in the period. The table is reproduced with permission from Hoelzemann et al. (2025).

Table E2: Quantitative Results of Experiment with AI Agents (Baseline)

Panel A: Round-level Outcomes

	Individual payoff	I(Individual found max)	I(Group found max)
	(1)	(2)	(3)
High	6.777***	0.022***	-0.001
	(0.183)	(0.007)	(0.002)
Low	0.067	-0.011	0.000
	(0.207)	(0.008)	(0.002)
Medium	-1.805***	-0.964***	-0.969***
	(0.182)	(0.008)	(0.017)
Constant	15.423***	0.978***	1.000***
	(0.176)	(0.007)	(0.001)
Round order FE	Yes	Yes	No
Payoff structure FE	Yes	Yes	Yes
Observations	2500	2500	500

Panel B: Analysis of Mechanisms

	Exploration	Individual payoff		I(Individual found max)		I(Group found max)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Round	Period 1	Period 2	Period 1	Period 2	Period 1	Period 2
High	-100.000***	6.569***	0.208***	0.802***	0.040***	0.010	0.000
	(0.085)	(0.183)	(0.049)	(0.018)	(0.009)	(0.010)	(0.002)
Low	-0.100	0.077	-0.009	0.007	-0.006	0.002	0.000
	(0.115)	(0.205)	(0.057)	(0.021)	(0.011)	(0.012)	(0.001)
Medium	-98.250***	2.258***	-4.063***	-0.196***	-0.946***	-0.980***	-0.970***
	(0.806)	(0.179)	(0.060)	(0.018)	(0.010)	(0.014)	(0.017)
Constant	100.000***	4.531***	10.892***	0.198***	0.960***	0.990***	1.000***
	(0.049)	(0.176)	(0.049)	(0.018)	(0.009)	(0.010)	(0.001)
Round order FE	No	Yes	Yes	Yes	Yes	No	No
Payoff structure FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	500	2500	2500	2500	2500	500	500

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. Estimates from OLS models. In Panel A, the sample in Columns 1 and 2 is at the participant-round level (5 participants × 500 rounds). The sample in Column 3 is at the group-round level (500 rounds). *Individual payoff*= participant-level round payoffs in Canadian dollars; *I(Individual found max)*:0/1=1 if the location of the maximum was found by the participant; *I(Group found max)*:0/1=1 if the location of the maximum was found by at least one participant in the round. The excluded category captured by the constant is the condition without data. See text for more details.

In Panel B, the sample in Column 1 is at the group-round level (500 rounds). The sample in Columns 2, 3, 4, 5 is at the participant-period level (5 participants × 500 periods of each type). The sample in Columns 6 and 7 is at the group-period level (500 periods of each type). *Exploration*= share of unknown mountains explored in the round; *Individual payoff*= participant-level period payoffs in Canadian dollars; *I(Individual found max)*:0/1=1 if the location of the maximum was found by the participant in the period; *I(Group found max)*:0/1=1 if the location of the maximum was found by at least one participant in the period.

Table E3: Comparing LLM-based Simulations to Direct Elicitation

				Mean group payoff (%)			Mean group breakthrough (%)		
Configuration	**Data**	**Rivalry**	**Ambiguity**	**Humans**	**AI Agents**	**Prediction**	**Humans**	**AI Agents**	**Prediction**
1	None	FALSE	FALSE	68	69	75	99	100	80
2	Low			72	70	62	100	100	80
3	Medium			58	61	75	46	3	80
4	High			98	100	95	100	100	100
5	None	TRUE	FALSE	-	38	45	-	100	40
6	Low			-	37	62	-	100	80
7	Medium			-	37	75	-	100	40
8	High			-	34	75	-	100	100
9	None	FALSE	TRUE	-	69	50	-	99	20
10	Low			-	69	62	-	100	40
11	Medium			-	62	62	-	5	40
12	High			-	100	75	-	100	100

Note: This table shows the results when we try asking GPT-4 to predict the outcomes of the experiments. Column 1 lists the specific prediction task. Column 2 lists the data condition. Columns 3 and 4 list which theoretical assumptions apply. Column 7 shows GPT-4's direct prediction for the average group earnings (as a percentage of the maximum possible earnings). This is compared with the outcomes derived using human subjects (Column 5) and using AI agents (Column 6). Similarly, Column 10 shows GPT-4's direct prediction for the mean likelihood of a breakthrough, which occurs when at least one group member finds the diamond. This is compared with the outcomes derived using human subjects (Column 8) and using AI agents (Column 9). In Rows 1-4, we show the predictions for the baseline replication. In Rows 5-8, we show the predictions for the extension where we introduce payoff rivalry. In Rows 5-8, we show the predictions for the extension where we introduce payoff ambiguity.

Appendix References

Agarwal, S., I. H. Laradji, L. Charlin, and C. Pal (2024): "LitLLM: A Toolkit for Scientific Literature Review," *arXiv preprint arXiv:2402.01788*.

Aher, G., R. I. Arriaga, and A. T. Kalai (2023): "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies," *arXiv preprint*.

Anderson, L. B., D. Kanneganti, M. B. Houk, R. H. Holm, and T. Smith (2023): "Generative AI as a tool for environmental health research translation," *Geohealth*, 7.

Argyle, L. P., E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate (2023): "Out of one, many: Using language models to simulate human samples," *Political Analysis*, 31, 337–351.

Ashokkumar, A., L. Hewitt, I. Ghezae, and R. Willer (2024): "Predicting Results of Social Science Experiments Using Large Language Models," *Working Paper*.

Bail, C. A. (2024): "Can Generative AI improve social science?" *Proceedings of the National Academy of Sciences*, 121.

Brand, J., A. Israeli, and D. Ngwe (2023): "Using GPT for market research," *Harvard Business School Marketing Unit Working Paper*.

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, et al. (2023): "Sparks of Artificial General Intelligence: Early experiments with GPT-4," *arXiv preprint*.

Carlson, N. and V. Burbano (2025): "The Use of LLMs to Annotate Data in Management Research: Guidelines and Warnings," *Available at SSRN: https://ssrn.com/abstract=4836620*.

Charness, G., B. Jabarian, and J. A. List (2023): "Generation next: Experimentation with AI," *NBER Working Paper w31679*.

Chong, D., J. Hong, and C. D. Manning (2022): "Detecting label errors by using pre-trained language models," *arXiv preprint arXiv:2205.12702*.

Csaszar, F. A. and D. A. Levinthal (2016): "Mental representation and the discovery of new strategies," *Strategic Management Journal*, 37, 2031–2049.

Dillion, D., N. Tandon, Y. Gu, and K. Gray (2023): "Can AI language models replace human participants?" *Trends in Cognitive Sciences*, 27, 597–600.

Doshi, A. R. and O. P. Hauser (2024): "Generative AI enhances individual creativity but reduces the collective diversity of novel content," *Science Advances*, 10, eadn5290.

Grimes, M., G. Von Krogh, S. Feuerriegel, F. Rink, and M. Gruber (2023): "From scarcity to abundance: Scholars and scholarship in an age of generative artificial intelligence," *Academy of Management Journal*, 66, 1617–1624.

Grossmann, I., M. Feinberg, D. C. Parker, N. A. Christakis, P. E. Tetlock, and W. A. Cunningham (2023): "AI and the transformation of social science research," *Science*, 380, 1108–1109.

Hagendorff, T. (2024): "Deception abilities emerged in large language models," *Proceedings of the National Academy of Sciences*, 121.

Hoelzemann, J., G. Manso, A. Nagaraj, and M. Tranchero (2025): "The streetlight effect in data-driven exploration," *NBER Working Paper w32401*.

Horton, J. and R. Horton (2024): "EDSL: Expected Parrot Domain Specific Language for AI Powered Social Science," Whitepaper, Expected Parrot.

Horton, J. J. (2023): "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" *NBER Working Paper w31122*.

Huang, J.-t., E. J. Li, M. H. Lam, T. Liang, W. Wang, et al. (2024): "How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments," *arXiv preprint*.

Kauffman, S. A. (1993): *The Origins of Order: Self-organization and Selection in Evolution*, Oxford University Press.

Kazemitabaar, M., R. Ye, X. Wang, A. Z. Henley, P. Denny, M. Craig, and T. Grossman (2024): "CodeAid: Evaluating a Classroom Deployment of an LLM-based Programming Assistant that Balances Student and Educator Needs," *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Lampinen, A. K., I. Dasgupta, S. C. Chan, H. R. Sheahan, A. Creswell, D. Kumaran, J. L. McClelland, and F. Hill (2024): "Language models, like humans, show content effects on reasoning tasks," *PNAS nexus*, 3.

Levinthal, D. A. (1997): "Adaptation on rugged landscapes," *Management Science*, 43, 934–950.

Li, P., N. Castelo, Z. Katona, and M. Sarvary (2024): "Frontiers: Determining the validity of large language models for automated perceptual analysis," *Marketing Science*, 43, 254–266.

Liang, W., Y. Zhang, Z. Wu, H. Lepp, W. Ji, X. Zhao, H. Cao, S. Liu, S. He, Z. Huang, et al. (2024): "Mapping the increasing use of llms in scientific papers," *arXiv preprint arXiv:2404.01268*.

Ma, P., R. Ding, S. Wang, S. Han, and D. Zhang (2023): "InsightPilot: An LLM-empowered automated data exploration system," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 346–352.

Manning, B. S., K. Zhu, and J. J. Horton (2024): "Automated social science: Language models as scientist and subjects," *NBER Working Paper w32381*.

Mei, Q., Y. Xie, W. Yuan, and M. O. Jackson (2024): "A Turing test of whether AI chatbots are behaviorally similar to humans," *Proceedings of the National Academy of Sciences*, 121.

Mohammadi, B. (2024): "Wait, It's All Token Noise? Always Has Been: Interpreting LLM Behavior Using Shapley Value," *arXiv preprint*.

Park, J. S., J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein (2023): "Generative Agents: Interactive Simulacra of Human Behavior," *arXiv preprint arXiv:2304.03442*.

Qu, L., S. Wu, H. Fei, L. Nie, and T.-S. Chua (2023): "LayoutLLM-T2I: Eliciting Layout Guidance from LLM for Text-to-Image Generation," *Proceedings of the 31st ACM International Conference on Multimedia*, 643–654.

Salvagno, M., F. S. Taccone, and A. G. Gerli (2023): "Can artificial intelligence help for scientific writing?" *Critical care*, 27, 75.

Si, C., D. Yang, and T. Hashimoto (2024): "Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers," *arXiv preprint*.

Tjuatja, L., V. Chen, S. T. Wu, A. Talwalkar, and G. Neubig (2023): "Do LLMs exhibit human-like response biases? A case study in survey design," *arXiv preprint*.

Wu, Y., X. Tang, T. M. Mitchell, and Y. Li (2023): "Smartplay: A benchmark for LLMs as intelligent agents," *arXiv preprint*.

Xu, Y., S. Wang, P. Li, F. Luo, X. Wang, W. Liu, and Y. Liu (2023): "Exploring large language models for communication games: An empirical study on werewolf," *arXiv preprint*.

Ye, Y., J. Hao, Y. Hou, Z. Wang, S. Xiao, Y. Luo, and W. Zeng (2024): "Generative ai for visualization: State of the art and future directions," *Visual Informatics*.