Strategy Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

How Well Can AI Do Strategy? Empirical Benchmarking Using Strategy Simulations

Ryan T. Allen, Rory M. McDonald

To cite this article:
Ryan T. Allen, Rory M. McDonald (2026) How Well Can AI Do Strategy? Empirical Benchmarking Using Strategy Simulations. Strategy Science 11(1):93-117. https://doi.org/10.1287/stsc.2025.0444

https://pubsonline.informs.org/journal/stsc

STRATEGY SCIENCE
Vol. 11, No. 1, March 2026, pp. 93–117
ISSN 2333-2050 (print), ISSN 2333-2077 (online)

How Well Can AI Do Strategy? Empirical Benchmarking Using Strategy Simulations

Ryan T. Allen,[a,*] **Rory M. McDonald**[b]

[a] Management Department, Marriott School of Business, Brigham Young University, Provo, Utah 84602; [b] Strategy, Ethics, and Entrepreneurship, Darden School of Business, University of Virginia, Charlottesville, Virginia 22903

*Corresponding author

Contact: ryan.allen@byu.edu, https://orcid.org/0000-0002-8227-8844 (RTA); mcdonaldr@darden.virginia.edu, https://orcid.org/0000-0003-4404-9212 (RMM)

Received: May 15, 2025
Revised: September 19, 2025; November 24, 2025
Accepted: December 15, 2025
Published Online in Articles in Advance: March 11, 2026

https://doi.org/10.1287/stsc.2025.0444

Abstract. Benchmarks have helped fuel rapid progress in large language models (LLMs) across a variety of domains including math, science, dialogue, and coding. Yet no existing benchmark adequately captures the defining elements of strategic decision making: uncertainty, complexity, irreversible multiperiod moves, and delayed or noisy feedback. This gap limits our ability to assess and guide LLMs' capabilities in strategy. We propose that established strategy teaching simulations provide an ideal benchmarking approach because (1) they approximate the essential features of real-world strategy, and (2) they do so in a controlled, replicable environment suitable for evaluation. To demonstrate this, we assess the performance of 21 proprietary and 13 open-source LLMs on the Back Bay Battery (BBB) simulation, a widely used exercise in strategy and innovation courses. The simulation requires balancing short-term profitability against long-term competitive positioning while integrating complex information about customer preferences and technological change. We built an interface enabling LLMs to interact with the simulation as though encountering it for the first time, masking identifiers to reduce contamination from prior training data. Our results show clear progress in composite BBB performance: Later models generally outperform earlier versions, and reasoning-focused models from late 2024–early 2025 (e.g., o4-mini, Claude Sonnet 4, Gemini 2.0 Flash) exceed even the average scores of historical MBA student cohorts. However, frontier models from mid-to-late 2025 (e.g., GPT-5, Claude Opus 4.5, Gemini 3) have declined, underperforming both earlier LLMs and MBA students. This decline is partially explained by a systematic bias toward exploiting the core business at the expense of investing in future growth. Overall, these findings highlight impressive advances in LLMs' strategic abilities since their inception. At the same time, we document current frontier models' surprising weakness in managing strategic uncertainty. This paper pioneers and provides guidance for using simulation-based benchmarking as a productive framework for strategy researchers to track progress, identify blind spots, and shape the trajectory of strategy-specific LLM capabilities.

History: Accepted for the Special Issue: Can AI Do Strategy?

Supplemental Material: The online appendix is available at https://doi.org/10.1287/stsc.2025.0444.

Keywords: decision making • information technology • artificial intelligence • technological change • technology strategy • competitive strategy

1. Introduction

How well can artificial intelligence (AI) make strategic decisions? Strategy scholars define such decisions as resource commitments and courses of action that shape long-term performance (Mintzberg et al. 1976, Eisenhardt and Zbaracki 1992, Csaszar 2018), such as entering a new market or allocating resources to long-term research and development (R&D) initiatives. They are unusually difficult because they are (1) complex and interdependent (Levinthal 1997, 2017; Rivkin 2000; Rivkin and Siggelkow 2006; Adner et al. 2014; Peterson and Wu 2021), (2) high stakes (Eisenhardt and Bourgeois 1988, Eisenhardt 1989), (3) involve multiperiod partially irreversible commitments (Ghemawat 1991), (4) yield delayed, noisy feedback (Csaszar 2018, Csaszar et al. 2024, Allen et al. 2026), and (5) occur under deep uncertainty in situations lacking clear precedents or rules (Mintzberg et al. 1976, Nickerson and Zenger

2004, Csaszar and Levinthal 2016, Felin and Zenger 2017, McDonald and Eisenhardt 2020, Allen 2025).

Supervised-learning AI systems developed in the last decade have excelled on tasks sharing *some* of these features—augmenting high-stakes decisions in radiology, information technology (IT) operations, human resources (HR) screening, and quantitative trading (Li 2017, Kleinberg et al. 2018, Arthur and Hossein 2019, Allen and Choudhury 2022, Csaszar et al. 2024); they have also surpassed human experts in complex strategy games like chess, Go, and StarCraft II (Newborn 2012, Silver et al. 2016, Vinyals et al. 2019, Gaessler and Piezunka 2023). Yet these capabilities have not readily translated to real-world strategy, at least in part because supervised learning depends on well-structured repeatable data (Choudhury et al. 2020, Csaszar et al. 2024, Felin and Holweg 2024), which are rarely, if ever, present in the context of actual strategic decisions (Eisenhardt 1989, Nickerson and Zenger 2004, McDonald and Eisenhardt 2020).

But the recent emergence of large language models (LLMs) has, for the first time, made it plausible for AI to carry out real-world strategic decisions. Unlike earlier supervised learning AI systems that were limited to prespecified tasks with structured data, LLMs can parse unstructured text and generate context-sensitive responses in a flexible way (Csaszar et al. 2024). This expands AI's capacity to engage in broader, unstructured reasoning processes that more closely resemble human cognition. As a result, AI has now surpassed various human benchmarks in coding, reading comprehension, multimodal reasoning, and even PhD-level science questions (Chen et al. 2021; Wei et al. 2022; Bubeck et al. 2023; Anthropic 2024; OpenAI 2024, 2025; Google 2025; Maslej et al. 2025).

Still, how well these enhanced capabilities extend to strategic decision making remains an open question. Early evidence shows GPT-3.5/4 can generate and evaluate business ideas comparably to human experts (Boussioux et al 2024, Csaszar et al. 2024, Doshi et al. 2025) and that using GPT-4 significantly enhances productivity in product development and consulting (Dell'Acqua et al. 2023, 2025). But these studies mostly assess one-shot, narrow tasks that sidestep key elements of strategic decision-making (i.e., uncertainty, multiperiod irreversibility, competitive dynamics, and delayed feedback). Meanwhile, recent conceptual work on AI cautions that LLMs' backward-looking, pattern-based reasoning may lack the forward-looking causal logics central to strategic value creation (Felin and Zenger 2017, Felin and Holweg 2024, Felin et al. 2024). Further complicating the question, prior empirical studies have typically examined human–LLM collaboration for a single model (e.g., GPT 3.5/4) at a time. Because of rapid changes in the underlying technology, standalone results for GPT-3.5/4 may not hold for later models, leaving the strategic capabilities of LLMs uncertain and state of strategy research on AI in a state of perpetual evolution as the underlying models improve.

These questions and issues persist in part because the field lacks standardized strategy-specific LLM benchmarks (Csaszar 2025). In other domains (e.g., coding, chat performance, PhD-level science, and video games), agreed upon LLM benchmarks have provided an objective yardstick for tracking various dimensions of LLM performance to inform the limits of their use in specific domains (Rein et al. 2024, Wang et al. 2024, White et al. 2024, ClaudePlaysPokemon 2025, UC Berkeley SkyLab 2025). Such benchmarks have significantly improved the quality of the systems they track, spurring on the respective fields by defining targets for further research and development. Meanwhile, LLM industry leaders have actively called for more domain-specific benchmarks to push forward model performance (OpenAI Pioneers Program 2025).

To both track and catalyze strategic decision-making capabilities, we propose a novel benchmarking approach that evaluates LLM performance on established strategy simulations. Many strategy simulations establish conditions conducive to benchmarks. They are designed to mirror the complexity and uncertainty of real-world strategic decisions while still providing a controlled, validated, and repeatable environment necessary for standardized benchmarking. Unlike reinforcement-learning achievements trained on millions of past plays (as in chess, StarCraft, and Go), we develop an approach for testing how LLMs perform when *encountering strategic decisions for the first time*, relying solely on general reasoning capabilities and without the aid of training, prior data, examples, or additional fine-tuning.

In this paper, we instantiate a strategy-specific benchmark using the widely used Back Bay Battery (BBB) simulation (Christensen and Shih 2019), where the decision maker allocates an R&D budget across an established core technology (≈80% of initial revenue) and a new emerging technology (≈20%) over eight periods. The simulation challenges the user (in this case, the LLM) to make decisions under uncertainty, navigate tradeoffs between short-term profitability and long-term competitive positioning, and synthesize a wide array of complex and dynamic information about customers, competitors, and evolving technologies. As in real-world strategy, there is not a single performance outcome; instead, multiple outcomes such as profitability, revenue, and growth in the new technology must be weighed and prioritized. Drawing on this insight, we develop a composite score that captures three key outcomes: cumulative profit (core business management proxy), cumulative revenue (overall expansion proxy), and final revenue from the emerging technology (future positioning proxy).

We ran hundreds of simulations over a range of leading proprietary (e.g., GPT-4o, o3, GPT-5, Claude Sonnet 4, Gemini 2.5 Pro, Grok 4) and open-source

(e.g., DeepSeek, Gemma, Llama, Qwen) LLM models. LLMs were able to interact directly with the simulation using a custom-built interface that delivered all simulation information directly to the LLM and then managed submitting LLM responses into the simulation decision-input fields. The interface ensured that each model interacted with the simulation consistently, without variable human intervention. It was also designed to mask the information sent to the LLM to prevent contamination from potential prior knowledge of data about the exercise, thereby presenting the situation as though the LLM was encountering it for the first time. Although our primary comparisons are model-to-model performance, we also compiled historical performance outcomes of 249 MBA students at a top U.S. East Coast business school to contextualize these results.

Generally, the results show marked improvement in LLMs' BBB composite scores over time, with early models like GPT-3.5 and GPT-4 performing substantially worse than later models. Yet, the best-performing models were not the latest models: OpenAI's o3-mini achieved the highest composite score, followed by o4-mini, Claude Sonnet 4, and Gemini 2.0 Flash. Released in early 2025, these models excelled at BBB by skillfully timing investments and pricing to achieve the outcome of both strong profitability and growth. As a result, these models outperformed the historical MBA student average score. By contrast, the current frontier models (including GPT-5, o3, and Gemini 2.5 Pro), released over six months later, performed considerably worse than the lighter models—and below the MBA student average. Although these current frontier models generated high cumulative profit, they concentrated heavily on the core existing technology, resulting in lower emerging-technology revenue growth.

Interestingly, although LLMs have consistently improved their performance in domains with public benchmarks over time,[1] their performance on BBB strategy initially followed this trend but has recently fallen away. For the latest models, the relationship between gains on other domain benchmarks and BBB strategy performance has plateaued—or even reversed. These breaks in trend may imply that advances in abilities such as general reasoning and conversational fluency—while correlated with stronger strategic outcomes up to a point—may no longer reliably predict them. Instead, the most recent models appear increasingly optimized for benchmarks in other domains, potentially at the expense of the forward-looking strategic reasoning that the BBB simulation captures. This disconnect underscores the necessity of strategy-specific benchmarks and provides a key motivation for our paper: Without such benchmarks, scholars and developers risk mistaking progress in adjacent domains for progress in strategic decision making. Moreover, the field lacks a shared infrastructure for systematically assessing and steering AI's strategic capabilities (Csaszar 2025). Strategy researchers have a pressing opportunity to develop, adopt, and refine strategy-focused benchmarks that reflect the distinctiveness of strategic decision making.

This paper makes three contributions to research on AI in strategy. First, the results show that some recent models appear to have crossed a key threshold in strategic cognition: Upon a first encounter with an open-ended uncertain strategy simulation, some LLMs can reason well enough to match or even surpass historical human manager simulation scores. This represents a notable expansion of AI's cognitive frontier, and hints at its capacity to reason effectively in novel, complex situations (Felin and Holweg 2024). Second, the study surfaces underappreciated weaknesses in frontier systems—most notably an apparent bias toward exploitative, backward-looking choices (Felin and Holweg 2024). The models' divergence in performance across benchmarks suggests that the microfoundations of successful strategic decision making may differ in important ways from those underlying success in other domains like coding or advanced science. Finally, our proposed replicable and extendable benchmarking framework enables strategy researchers and developers to track and guide the creation of systems with stronger performance in contexts that bear the hallmarks of strategic decision making.[2] We end with concrete guidance and principles for future strategy-specific benchmarking efforts in the field, positioning our framework as a practical guide for subsequent work on AI benchmarks in strategy.

2. Literature Background

Our focal question of whether current frontier AI can make high-quality strategic decisions is informed by two prior bodies of research documenting: (1) rapid improvements in large language models (LLMs) on general cognitive benchmarks and (2) emerging evidence of such models augmenting strategy-related tasks. A review of these literatures points to a key opportunity for developing a suitable benchmark for tracking AI's progress in strategic decision making.

2.1. Rapid AI Advances in General Performance Benchmarks

Recent LLMs have rapidly improved across benchmarks in a range of domains. For instance, progression from GPT-3 to GPT-4 marked significant advances in coding, with pass rates on the HumanEval benchmark (assessing Python programming ability) rising from around 28% for early GPT-3 codex models to 67% for GPT-4 (OpenAI et al. 2023). Similar improvements occurred in reasoning and reading comprehension,

with GPT-4 achieving 86.4% accuracy on the massive multitask language understanding (MMLU) benchmark, a notable increase from GPT-3.5's approximately 70% and approaching human expert levels (OpenAI et al. 2023). GPT-4 has also approached or exceeded human-level performance in professional knowledge tests, scoring in the 90th percentile on the Uniform Bar Exam—surpassing most law graduates and vastly outperforming GPT-3.5's bottom 10th percentile (OpenAI et al. 2023). In medical licensing exams, GPT-4 achieved approximately 80% accuracy (e.g., MedQA), comfortably exceeding GPT-3.5's ~58%, and approaching the performance of medical school graduates (OpenAI et al. 2023). This research illustrates that LLMs have rapidly transitioned in a relatively short period of time from foundational competencies to expert-level performance in complex cognitive tasks.

In 2024 and 2025, OpenAI's "thinking" models using "chain of thought" reasoning—such as o1 and o3—have continued to rapidly advance (Wei et al. 2022). They have achieved remarkable results across a range of benchmarks that were not possible with the base language models like GPT-4. The o1 model, introduced in late 2024, ranked in the 89th percentile on Codeforces and solved 83% of International Mathematics Olympiad qualifier problems, a significant improvement over GPT-4o's 13% (OpenAI 2024). The subsequent o3 model achieved 96.7% accuracy on the American Invitational Mathematics Examination (AIME) and 87.7% on graduate-level science exams (OpenAI 2025). Comparable performances have been reported by Anthropic's Claude 3 and Google's Gemini 2.5 models (Anthropic 2024, Google 2025).

One of the frontier models as of August 2025 is the GPT-5 model from OpenAI, which has achieved top (or near top) performance across a range of benchmarks including AIME (high-school math questions), GPQA (PhD-level science questions), LiveBench (a range of cognitive tests), and LMArena (testing chat performance) (White et al. 2024, UC Berkeley SkyLab 2025). Overall, such progress indicates the rapid and substantial expansion of AI capabilities into areas previously exclusive to highly trained human experts (Chen et al. 2021, Bubeck et al. 2023).

These leaps, particularly in the recent "thinking" models' performance on high-level expert tasks like math and science, suggest not just incremental improvements but a fundamental shift in AI's problem-solving capabilities (Wei et al. 2022). Such rapid progression from basic proficiency to surpassing human performance in complex domains underscores the accelerating trajectory of AI development and its expanding potential to tackle more tasks once considered exclusive to expert human intelligence.

2.2. AI Performance in Strategic Decision Making

To what extent do these rapidly advancing AI capabilities apply to strategic decision making? Empirical evidence suggests that even early models like GPT-3.5 and GPT-4 could replicate human performance in generating and evaluating business ideas. For example, evaluators consistently preferred GPT-3.5–edited business plans over human-generated versions by 0.14 standard deviations (Csaszar et al. 2024). In related research, GPT-3.5's evaluations of startup strategies correlated ($r = 0.52$) with expert judges' scores, indicating comparable evaluation quality (Csaszar et al. 2024). In another study, AI business ideas were rated by judges as higher quality than human ideas, despite the idea generation costing much less than human crowds (Boussioux et al. 2024).

Further research using GPT-4 explored the reliability of AI evaluations by generating 60 startup business models and having them assessed in pairwise comparisons by ensembles of LLMs, which aligned closely with evaluations from human strategy experts (Doshi et al. 2025). A follow-up experiment validated these findings with 60 real-world business plans from a U.S. university startup competition (Doshi et al. 2025). Although the LLMs often produced evaluations that were inconsistent and biased, in aggregate, the rankings tended to resemble those of human experts. These studies collectively suggest that aggregating LLM judgments across multiple models, roles, and prompts can produce early startup pitch assessments comparable to those of human experts.

Field experiments have further demonstrated AI's potential to enhance human productivity in strategy-related organizational settings. A study involving 776 Procter & Gamble professionals randomly assigned participants to product innovation challenges, with or without GPT-4-powered generative AI assistance (Dell'Acqua et al. 2025). Independent expert evaluations showed that AI significantly boosted solution quality by 0.37 standard deviations, alongside improvements in novelty and feasibility. Similarly, a field experiment with 758 consultants at Boston Consulting Group examined GPT-4's impact on knowledge work tasks (Dell'Acqua et al. 2023). AI significantly enhanced performance on a range of tasks—improving consultant output quality by more than 40%, increasing task completion rates by 12.2% and reducing task completion times by more than 25%.

However, evidence from the Dell'Acqua et al. (2023) study also suggest a "jagged technological frontier" where GPT-4 performed worse on complex, context-heavy tasks that required synthesizing noisy, firm-specific data. In the field experiment, BCG consultants using GPT-4 were 19 percentage points less likely to reach the correct recommendation on an out-of-frontier business case. They finished faster and earned

higher recommendation-quality scores for their write-ups, yet these results underscore the risk of confidently wrong outputs on the tasks that most closely resembled truly strategic decisions.

Other scholars have raised conceptual critiques of AI's ability to make valuable strategic decisions. Felin and Holweg (2024) caution that "AI uses a probability-based approach to knowledge and is largely backward-looking and imitative, while human cognition is forward-looking and capable of generating genuine novelty." According to these arguments, LLMs' pattern-driven, retrospective reasoning may lack the theory-based, forward-looking causal logic at the core of creating value in strategic decision making (Felin and Zenger 2017, Felin et al. 2024). These critiques implicitly question the extent to which early evidence of AI's positive impact in strategy-related tasks can inform our understanding of LLM performance in actual strategic decisions.

Together, these studies establish a useful baseline in demonstrating that AI, even in relatively early generations like GPT-4, can mimic human-level performance in strategy-related tasks like generating and evaluating early business plans. However, LLMs appear to excel in some domains more than others, and early evidence and conceptual critiques suggest that they may not be particularly well suited to decisions that bear the hallmark of truly strategic situations (e.g., complexity, uncertainty, noisy feedback, etc.).

2.3. Need for a Strategic Decision-Making Benchmark

Although valuable, existing strategy-adjacent benchmarks and empirical studies on AI in strategy provide only limited insight into LLMs' capacity for real-world strategic decision making (Csaszar 2025). Table 1 summarizes the shortcomings of prior benchmarks and empirical tests. For example, although games such as chess, Go, and StarCraft contain many hallmarks of strategy, the remarkable performance of supervised-learning systems in these domains depends on training with millions of games and well-defined outcomes—an approach that neither applies to LLMs nor translates to real-world strategic contexts. Other LLM-based game benchmarks, such as *Claude Plays Pokémon*, avoid these extensive training requirements but lack the complexity, uncertainty, and irreversibility that characterize business strategy. Similarly, studies that evaluate LLMs' ability to generate or assess business ideas capture narrow, isolated tasks rather than the multiperiod and path-dependent nature of strategy. Compounding the problem, empirical studies risk quickly becoming outdated as models evolve, reducing their long-term relevance for both scholars and practitioners. In sum, despite the value of these early efforts to study AI and strategy, their limitations constrain our ability to meaningfully assess LLMs' strategic capabilities.

These shortcomings underscore the need for a domain-specific strategy benchmark for LLMs. A strategy-oriented benchmark—analogous to those established in coding, mathematics, and games—would provide a stable and replicable foundation for evaluating performance, guiding model deployment, and preserving the interpretability of prior research findings. Unlike narrow tasks such as rating business plans or generating product ideas, established business simulations more closely approximate the conditions of real-world strategy: They can embed complexity, multiperiod decision making, incomplete and noisy feedback, and significant uncertainty. At the same time, because they are standardized and controlled from the researcher's perspective, simulations provide a replicable environment suitable for long-term benchmarking. In our proposed approach, these simulations can be presented as novel decision problems, encountered without prior examples or fine-tuning, thereby preserving their value as a never-before-encountered strategic issue from the perspective of the LLM.

Our proposed strategy-specific benchmark thus fills a critical gap (Csaszar 2025). By embedding the distinctive features of strategic choice in a controlled, repeatable environment, such a benchmark enables the systematic evaluation of LLMs' unaided decision-making capabilities and allows findings to accumulate over time. Just as benchmarks in coding and reasoning have accelerated progress in those domains, a strategy benchmark has the potential to not only measure but also spur improvements in the models by clearly highlighting the strengths and weaknesses of LLMs in strategic decision-making across a range of models over time.

3. Benchmarking Methodology

3.1. BBB Simulation Overview

In this paper, we use the BBB simulation, a widely adopted software tool in business education, as a first step in benchmarking LLMs' strategic decision-making capabilities (Christensen and Shih 2019). In the simulation, participants act as general managers of a fictional battery manufacturer over an eight-year period. Each round, they allocate a limited R&D budget across potential technology investments (e.g., energy density, recharge time, process improvement), set prices, and forecast sales for two technologies: the core Absorbed Glass Mat (AGM) business, which initially generates 80% of revenue, and the emerging Supercapacitor (SC), which initially accounts for 20%. Participants must also manage three distinct customer segments (Automobiles, Warehouses, and Uninterruptible Power Supplies (UPS)) each with their own technology performance dimension priorities.

Table 1. Existing Evidence of AI Strategy Capabilities

Feature	Chess (CCRL), Go (CGOS), StarCraft II (AI Arena)	LLM video games (e.g., Claude Plays Pokémon)	Idea generation and evaluation (Csaszar et al. 2024, Doshi et al. 2025)	Back Bay Battery (this paper)	Future strategy benchmarks
Elements of strategy					
Complexity	✓	○	○	✓	
Multiperiod dynamics	✓	✓	—	✓	
Irreversible commitments	✓	○	NA	✓	
Delayed/noisy feedback	○	—	NA	✓	
Uncertainty/novelty	—	—	○	✓	
Business context					
Business stage	NA	NA	Early pitch	Strategic resource allocation under uncertainty	Other business contexts and stages
Primary outcome	Elo/win rate	Game completion/ score	Idea or evaluation quality	Revenue growth, profitability, survival	Other outcomes like ethical considerations; external stakeholder buy-in, etc.
Interaction structure					
Multiagent competitive interaction	✓	—	—	— [a]	Multiplayer simulations: model versus model and model versus human
Agentic Abilities	NA	✓	—	—	Play in same interface as human; test ability to manage in real-time
AI characteristics					
LLM	—	✓	✓	✓	
Zero-shot	—	✓	✓	✓	Training with past relevant cases and data

Notes. Rows are ordered by prevalence of ✓ across benchmarks. ✓, present; ○, partially present; —, absent; NA, not applicable; Zero-shot, inference-only (no highly computational task-specific training or fine-tuning).
[a]Back Bay simulates external competition with news events and downward pricing pressure, but it is not truly a multi-agent game.

The simulation is designed to capture the classic strategic dilemma of balancing exploration and exploitation over time (March 1991, Benner and Tushman 2003, Gupta et al. 2006). The core technology (AGM) provides reliable cash flows, but there are indications that it will face eventual obsolescence and market saturation. In contrast, the new technology (SC) represents an unproven but potentially high value emerging growth market. It begins as a niche solution with limited market appeal but has the potential to redefine market boundaries and become a major growth driver if developed to meet emerging customer needs (Christensen and Bower 1996, Adner 2002, Adner and Zemsky 2005, Christensen et al. 2018). Early investments in SC, however, do not yield immediate results. Therefore, R&D investment in the new technology reduces short-term profits; successful growth requires strategic foresight, calculated risk taking, and sensitivity to evolving market signals.

In the simulation, participants face significant complexity: Market dynamics evolve based on simulated customer preferences, competitor behavior, and technological progress, none of which follow simple or easily predictable patterns. The interdependent decisions carry multiyear consequences, because neglecting one technology or overcommitting to another can erode long-term competitiveness (Peterson and Wu 2021). Feedback is delayed and noisy, because the effectiveness of R&D investments may only become apparent several years (decision-periods) after they are made. Success in the simulation requires not only logical reasoning but also the ability to synthesize a deluge of possibly relevant information and make high-stakes decisions under significant uncertainty.

In the primary benchmarking exercise, we use the "advanced" difficulty in "legacy" mode, which is the standard in MBA classrooms and matches our

historical MBA student comparison set. In the advanced difficulty mode, participants manually enter sales forecasts; these forecasts determine projected revenue and cap R&D at 3% of that value. In this mode, participants can be "fired," triggering an end to the simulation, if their sales variance ((actual unit sales − projected unit sales)/projected unit sales) reaches −50% in any single year, remains below −20% for more than three years, or if the division posts negative net income for three consecutive years. In the sensitivity checks below, we also show similar results using the "basic" difficulty mode, which removes the possibility of being fired. Although deterministic (ensuring repeatability and consistency), from the perspective of the first-time user, the simulation offers sufficient complexity to map onto participants' decision-making capabilities in an environment that resembles a novel and challenging strategic situation (Christensen and Shih 2019).

3.1.1. Simulation Steps. Participants begin the simulation with a briefing that outlines the firm's market position, customer segments, and technologies (see Online Appendix 1, step 1). This informs them that as general manager of BBB, they are tasked with managing two distinct technologies (AGM and SC) and serving three commercial markets: Automobiles, Warehouse Equipment, and UPS systems. After reading the briefing, participants process information and make decisions across two main tabs:

1. **Analysis Tab:** Participants review detailed performance dashboards covering sales, profitability, customer preferences, technological performance, and market news updates (see Online Appendix 1, steps 2 and 3).
2. **Decision Tab:** For each technology (AGM and SC), participants set sales forecasts, unit prices, and allocate R&D budgets for the year across five technology performance dimensions (see Online Appendix 1, step 4).

Once the year's decisions are submitted, the simulation processes the results, updating market demand, technological progress, and financial outcomes based on the participant's choices. These results then become the basis for the following year's decisions. This cycle repeats up to eight years, at which point the user either successfully completes the simulation or is fired for poor performance (see above for "firing" criterion).

3.1.2. Simulation Structure and Dynamics. The BBB simulation models the challenges of managing an established business (AGM batteries) while deciding whether and how to invest in an emerging disruptive technology (SC). The simulation creates tension by limiting resources: There is never enough R&D budget to fund all opportunities. A common failure mode is spreading investments too thinly across multiple initiatives, which prevents achieving meaningful progress in any area. Success requires carefully chosen tradeoffs, prioritization, and long-term strategic thinking.

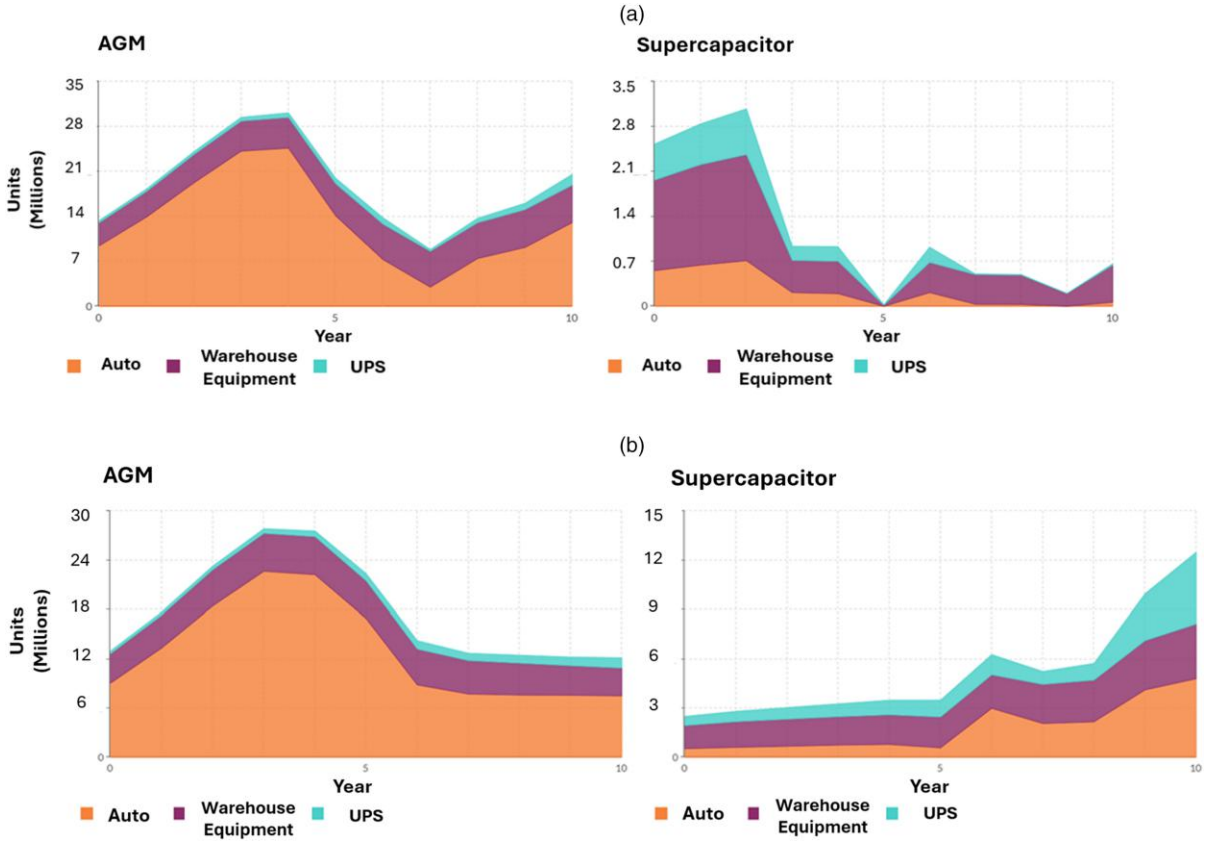
Several programmed dynamics drive the competitive environment. Around year 4, the simulation introduces a downturn in AGM sales due to increased foreign pricing pressure. This decline is largely unavoidable; even strong management choices cannot eliminate it entirely. However, by investing in process improvements (which lower costs) and tactically reducing prices, players can soften the impact, retain some market share, and sustain profitability in the core AGM business.

Meanwhile, investments in SC follow a cumulative, path-dependent trajectory. Early and sustained R&D is necessary—single exploratory investments almost never pay off. SC breakthroughs take multiple years of funding in focused performance dimensions that underperform customer demands (such as energy density or cycle life) before they begin to deliver tangible market demand. Importantly, these investments are risky: Players commit capital without guaranteed success, only knowing that if breakthroughs occur, SC can eventually provide a far superior value proposition for select customer segments compared with AGM. When a key technological threshold is crossed, the SC market can accelerate rapidly, creating the potential for exponential growth.

3.1.3. Illustrative Paths to High Performance. One viable route to high performance is to focus entirely on the established AGM technology. In this scenario, players concentrate most R&D dollars on AGM process improvements (to reduce cost and improve margins). This strategy can lead to strong cumulative profit scores because of steadily improving margins in the core business. However, it has a structural limitation: Because SC is ignored, there is no opportunity for breakout growth as the core business reaches obsolescence and market saturation. By the later years of the simulation, even though the company may have achieved high levels of profit, the company is typically weaker overall than at its starting point. This outcome is illustrated in Figure 1(a).

The highest performance usually comes from balancing near-term AGM profitability with early bets on SC. In this approach, players initially invest in AGM process improvements to maintain competitiveness and cash flow. Simultaneously, they make early but modest R&D investments in SC dimensions most demanded by customers (commonly energy density and cost/process efficiency). When the inevitable AGM downturn arrives, managers use price adjustments and process efficiency to mitigate losses, whereas their cumulative SC investments begin to mature. If the timing is right, the company can pivot into SC just as the technology becomes ready, riding the wave of its explosive growth. This balanced approach is shown in Figure 1(b).

Figure 1. (Color online) Illustrative Back Bay Battery Results



Notes. (a) Strong focus on AGM (sustaining/core technology) investment, while ignoring supercapacitor (emerging technology). (b) Successful transition to supercapacitor (emerging technology) while maintaining AGM (sustaining/core technology) steady state. Note different *y*-axis scales.

The simulation offers more opportunities to fail than to succeed. Failures can be tactical missteps: For example, managers may be fired for missing sales forecasts too often or for incurring large short-term losses—for instance, destroying demand with prices too high or removing margins with prices too low. Failures can also stem from strategic misalignment: A commonly occurring trap is spreading resources too thin, or over-committing to the new SC technology too early, which starves the firm of AGM profits needed to sustain investment. Without a cash-generating core business, players are unable to fund the multiyear SC effort, and bankruptcy or dismissal follows.

In sum, the BBB simulation highlights the classic innovator's dilemma: balancing exploitation of a profitable but declining core business with exploration of a risky, uncertain disruptive technology. Although there are many possible tactical variations, the structural lesson is consistent: Spreading investments too thin leads to mediocrity, ignoring disruption leads to decline, and over-committing too early leads to collapse (Christensen et al. 2018, Peterson and Wu 2021, Chu et al. 2025). Strategic balance, patience, and disciplined investment create the conditions for long-term success.

3.1.4. Simulation Outcomes. As in real-world strategic decision making, there is no single given simulation metric that defines optimal performance. In the simulation, participants can track various outcomes such as revenue, growth trends, and profitability across both technologies over time. To mimic real-world strategic decision making, students playing the simulation are not explicitly told whether to maximize profit, revenue, or growth in the new technology, each of which requires different priorities and tradeoffs to maximize (Christensen and Shih 2019).

In our benchmarking, we use three metrics that capture different aspects of performance: *Cumulative Profitability*, *Cumulative Revenue*, and *Emerging Tech Revenue*. *Cumulative Profitability* reflects total net income generated by both technologies across all customer segments over the eight simulated years. This serves as a proxy for effective management of the core business, including pricing, investment in the core technology and cost control. High profitability indicates maintaining a competitive edge in the core business while avoiding overspending. *Cumulative Revenue* reflects the total revenue generated by both technologies across all customer segments over the simulation

period, serving as a proxy for overall business expansion. *Emerging Tech Revenue* measures the amount of revenue from the emerging SC technology in the final year of the simulation. This serves as a proxy for well-timed investment in the new technology, positioning the company for future success and supporting a higher hypothetical valuation in the final year. To reduce correlation between the metrics, we subtract the *Emerging Tech Revenue* (i.e., the final period SC revenue) from the cumulative revenue total, so that cumulative revenue includes all revenue from both technologies in all periods except for the final period SC revenue.

These three outcomes highlight distinct aspects of performance in the simulation. In our data set of human participants (discussed below), cumulative profit and final emerging-technology revenue are weakly negatively correlated (−0.07), cumulative profit and cumulative revenue are weakly correlated (0.30), and cumulative revenue and final SC revenue show a moderate correlation (0.56). In our LLM data set, the correlation between cumulative profit and final emerging-technology revenue is 0.20, between cumulative revenue and final emerging-technology revenue is 0.763, and between cumulative profit and cumulative revenue is 0.508. Together, these measures provide a multidimensional view of strategic performance, capturing financial outcomes, market positioning, and responsiveness to emerging opportunities.

To simplify comparisons, we created a single summary benchmark measure, which we call the BBB *Composite Score*. Following the precedent of other benchmarks (Wang et al. 2024, White et al. 2024, Open LLM Leaderboard 2025), the score is calculated as the evenly weighted, normalized performance across the three core outcomes: *Cumulative Profitability*, *Cumulative Revenue*, and *Emerging Tech Revenue*. We normalize each outcome using min–max scaling, then combine the results with equal one-third weights. The hypothetical best score is one (the top-performing run across all three metrics), and the worst is zero (the lowest-performing run across all three metrics). It is important to emphasize that this within-data set normalization technique implies that the *Composite Score* is useful only for comparisons within the same data set, and it does not have intrinsic meaning when comparing across data sets.[3] The score is also sensitive to the specific weighting scheme applied. For these reasons, it is intended as a simplifying visualization rather than a standalone metric, and it should always be interpreted alongside the underlying outcome measures.

3.2. Human Benchmarking Data

Although the primary purpose of our study is to compare performance across AI models, we also collected historical human performance data to contextualize the AI results. We collected historical performance on the simulation from second-year MBA students enrolled in a strategic innovation course at a top U.S. university. The human benchmarking data set comprises first-run simulation results from 249 students across five years (2018, 2020, 2023, 2024, and 2025).[4] Students completed the BBB simulation individually, under "advanced" difficulty settings in "legacy" mode (the same as used in the LLM benchmarking). Importantly, we consider only each student's first full run, ensuring that outcomes reflect decision making under initial uncertainty rather than the benefits of repetition or hindsight.

The data set reflects hundreds of hours of work by highly skilled participants. Each student invested approximately one to two hours in completing the simulation. Given the selectivity of the university's MBA program and the extensive strategic training these students receive (typically toward the end of their second year), this benchmarking population represents a high standard of human capital and business expertise. Students were enrolled in a course covering such topics as exploration, exploitation, and disruption, which likely primed them to expect such dynamics in the simulation. They completed the exercise on their own time, outside of class. They were explicitly instructed not to use outside materials, and they had no explicit incentive (grade or otherwise) to cheat.

Because of course topics priming and the uncontrolled (completed outside of class) environment, the MBA comparison is imperfect and is not intended as the primary comparison of our study. Nevertheless, it provides valuable context for evaluating model performance against how highly trained human strategists perform in the same simulation.

To ensure robustness and generalizability of the human performance data, we conducted checks for potential cohort and temporal biases. We did not find a significant change in cumulative profitability across years (see Appendix A). Although cumulative revenue was modestly higher in the earlier cohorts (2018 and 2020), overall benchmarking relative to the AI results yields nearly identical conclusions, regardless of which cohorts are used.

These checks were important for two reasons. First, the simulation software underwent an update between 2020 and 2023 that changed the labeling of the two core technologies, although all later cohorts used a version compatible with the earlier simulation dynamics, which allowed for comparability between the years. Second, although students in 2023–2025 were explicitly instructed not to use AI tools like ChatGPT, we cannot rule out the possibility that some may have done so—making it essential to verify that later cohorts do not significantly differ from earlier ones. Additional robustness analyses, including re-estimations on

different data subsets, confirmed that our core findings hold consistently across all cohorts.

3.3. AI-Simulation Interface

3.3.1. Overview. We developed a program enabling LLMs to interact with the BBB simulation, without prior training or simulation-specific knowledge. Using automated web browsing (Selenium), our system launches the simulation, retrieves relevant data to send to the LLM each round, and submits the resulting LLM-generated decisions. For pseudocode that provides a simple overview of this process, see Appendix B.

In each simulated year, the program first collects the same information available to human players in the "Analyze" tab, including sales data, customer preferences, R&D results, financial performance, and market news. It also retrieves the current R&D budget constraint, which limits how much the player can invest in new technologies. These data are then reformatted and passed to the model as input—along with the simulation overview provided on the simulation landing page and a clear formatting prompt.

The program interacts with the LLMs using API calls via LiteLLM, a model-agnostic client. To mitigate model drift, we pin models to versioned API snapshots wherever the provider supports them (e.g., OpenAI model "snapshots" and Anthropic's date-suffixed model IDs), which according to their documentation, locks behavior to a specific release rather than an auto-updating alias. In contrast, consumer chat apps (e.g., ChatGPT) are routinely and silently updated to new model builds. For providers and configurations where our strings referenced rolling aliases (e.g., Google Gemini family names and some xAI Grok family names), we report the exact alias and run dates, and we executed all comparisons within the same window. See Appendix C for the model names used and the dates they were run.

In each simulated year, the model is fed all the information from the simulation and is instructed to respond with a list of 14 numbers (sales forecasts, pricing decisions, and R&D allocations across the two technologies and five R&D categories) and a textual explanation of the reasoning behind their strategy. The responses are stored, and the numbers are parsed and checked to ensure they follow the correct format and stay within budget. If not, the model is prompted to revise its decisions, with up to three retries allowed. Once a valid set of decisions is obtained, they are entered back into the simulation interface, just as a human user would do manually.

After submitting the decisions, the program checks whether the simulation has ended. If not, it proceeds to the next year. At the end of each run, we log the model's performance outcomes, such as cumulative profit, cumulative revenue, and emerging tech revenue, and reset the simulation before beginning the next run. This automated loop allows us to run hundreds of simulations across different models, all under standardized conditions.

3.3.2. Masking the Input to Avoid "Preknowledge" Contamination. Contamination from "preknowledge" is a risk that top LLM benchmarking studies must mitigate (White et al. 2024). The key risk in this case is that LLMs were trained on a data set that included the BBB case (and derivative write-ups and solutions), so when prompted they can simply recall and implement the "successful playbook." Our research design would then merely test whether an LLM can reproduce a high-scoring policy by recalling its training data rather than by strategic reasoning about the decision environment.

To address this concern, our setup deliberately masks all text inputs to ensure that the LLM encounters what appears to be a never-before-encountered case. Before any text from the simulation is sent to an LLM, we programmatically transform it using a deterministic masking function. This function applies a series of regular-expression substitutions to all strings retrieved from the simulation—including the initial briefing, all on-screen labels, news items, tabular data, and performance metrics. Concretely, canonical identifiers that uniquely mark the simulation (e.g., "Back Bay Battery," "AGM," "Supercapacitor," "UPS," "Warehouse Equipment," and the associated performance dimensions such as "Energy Density" and "Recharge Time") are systematically replaced with synthetic names (e.g., "EnergyCo," "AeroBond Matrix," "Quantum Storage Cell," "Continuity Power Modules," "Industrial Handling Systems," "Specific Energy Index," "Recovery Time Index"). The same renaming logic is applied consistently across all periods and runs (see Appendix D for full regular expressions). As a result, the model never sees the original case name, technology labels, or segment names that appear in teaching materials and online descriptions of the BBB simulation. As an added precaution, we rewrote the initial "Background" brief submitted to the LLM to present the same substantive information but with entirely different wording, thereby avoiding the risk that models might be triggered by verbatim overlap with known BBB materials (see Appendix D for alternative wording).

By masking the terms from the simulation, we are not attempting to prevent the model from using generic strategic concepts learned from other sources (e.g., theories of disruption or exploration versus exploitation), because leveraging such general knowledge is precisely what we aim to evaluate. Our masking procedure is instead designed to remove the specific lexical and structural cues that would allow

the model to recognize this simulation as BBB and simply retrieve a rehearsed solution. Under standard assumptions about how LLMs use pretraining data, contamination of this benchmark would require that the model both (a) encountered detailed descriptions or solutions to the BBB case during training and (b) recognize our simulation as the same case at inference time, using distinctive lexical cues (e.g., exact names of the firm, technologies, and customer segments). Our masking procedure directly targets this recognition step. By renaming all case-specific entities and performance dimensions with synthetic labels that do not occur in the original materials, and by rewriting the background text, we remove the unique "fingerprints" that would license direct retrieval of a memorized solution. What remains available to the model are only general-purpose strategic concepts (e.g., how to balance investment in an established versus an emerging technology), which are precisely the capabilities we intend to test.

As a proof-of-concept validation of our masking protocol, we conducted a simple two-condition prompt test with six models (o3-mini, gpt-4o, gpt-5, Claude Sonnet 4, Gemini 2.5 flash, and Gemini 2.5 pro). In the first condition, we asked the model to "explain the Back Bay Battery simulation and how to win it" without any additional context. Four of six models produced detailed and consistent descriptions that matched public teaching materials, consistent with contamination. In the second condition, we instead asked about the "EnergyCo" simulation; under these conditions, no models were able to give helpful or relevant advice to the BBB simulation (four hallucinated fake simulations; two admitted no knowledge; see results in Online Appendix 2). These results illustrate that, although the model likely memorized aspects of the playbook from the original BBB case, the masked version does not trigger the same stored representation, supporting our claim that masking the specific lexical terms used in the BBB simulation effectively neutralizes direct recall of BBB materials.

Although no masking scheme can guarantee the complete absence of any indirect prior exposure, it is important to note that our inferences, findings, and comparisons below are about *relative* performance across models *under the same masked representation of the task*. Taken together, these checks suggest that masking succeeds in preventing direct playbook recall and that any potential leaks would affect levels of performance but likely not the relative differences we measure.

4. Benchmarking Results [Database]

4.1. Primary Benchmarking Results

Figure 2 displays simulation outcomes for the *Composite Score* (the combined performance of *Cumulative Profit*, *Cumulative Revenue*, and *Emerging Tech Revenue*) across 21 of the key flagship models from the top proprietary LLM providers, Anthropic, Gemini (Google), OpenAI, and xAI, according to each model's API release date. Figure 2 shows that, by and large, LLM performance has improved significantly across model generations. The early models from 2022 to 2023, particularly GPT-3.5-turbo, consistently perform below other models. They also perform substantially worse than the MBA student benchmark. Later, early-2025 flagship models like gpt-4o, claude-sonnet-4, gemini-2.0-flash, o4-mini, and o3-mini are significantly higher performing than the early models (from 2022 to 2023) and higher than the historical human MBA average, representing a significant achievement in the ability to navigate the strategic complexities of the simulation. Reading through the reasoning behind the LLM strategies, common responses include statements such as "The strategy prioritizes defending the core [AGM] business while making measured progress with [SC]" (Gemini 2.0-flash response).

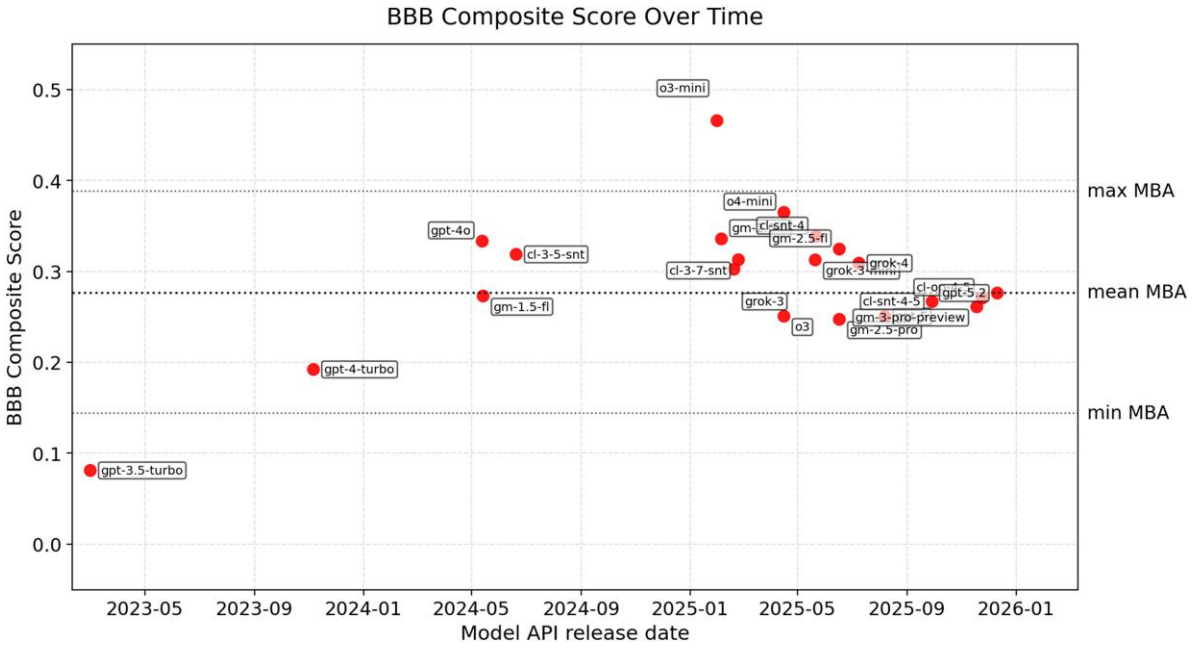
Surprisingly, however, the most recent frontier models from mid-2025 (Gemini-2.5-pro, o3, gpt-5) perform substantially worse than the early-2025 models and worse than the historical MBA average performance. Common reasoning included statements like "Pause all QSC R&D due to long lead times and poor fit with current market requirements" (GPT-5) and "We will not invest any R&D into QSC this year, preserving our limited capital for the core business" (Gemini-2.5-Pro).

Figure 3 shows the same results ranked by performance, along with raw simulation run–level results and confidence intervals. To better understand why the current frontier flagship models performed worse, Figure 4 breaks out separate scores for *Cumulative Profit*, *Cumulative Revenue*, and *Emerging Tech Revenue*. The figure clearly shows that, although models like GPT-5, o3, and Gemini-2.5-pro tended to achieve relatively high *Cumulative Profit*, they achieved very low *Cumulative Revenue* and *Emerging Tech Revenue*.

Figure 4 also displays the *Years Survived* in the simulation, indicating whether the simulation was completed (full eight years) or the run ended by being fired (for heavy losses or consistently missing sales forecasts). This indicates that the low *Cumulative Revenue* score of the frontier models was partially explained by the tendency to get fired earlier. However, even in years where they survived, the cumulative revenue achieved was relatively low compared with their cumulative profit. In sensitivity checks below, we show that the frontier models still fall into the same trap of over-focusing on profit while achieving lower revenue even in basic simulation mode, where it is not possible to be fired.

Figures 5 and 6 more clearly demonstrate a break in trend for LLMs on the BBB composite score versus

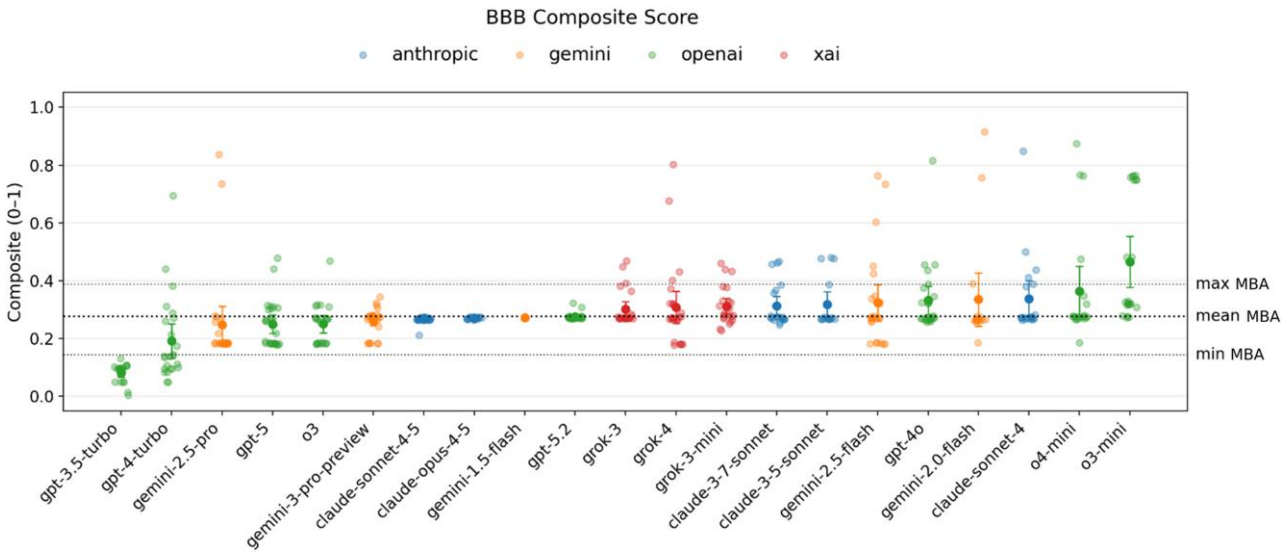
Figure 2. (Color online) Back Bay Battery Simulation LLM Performance: Composite Scores over Time



Notes. Each point indicates the per-model mean composite score. Dotted horizontal lines mark the historical MBA minimum, mean, and maximum composite scores for reference. Because the MBA results were collected under different conditions, they should be treated as contextual reference points rather than a direct apples-to-apples comparison.

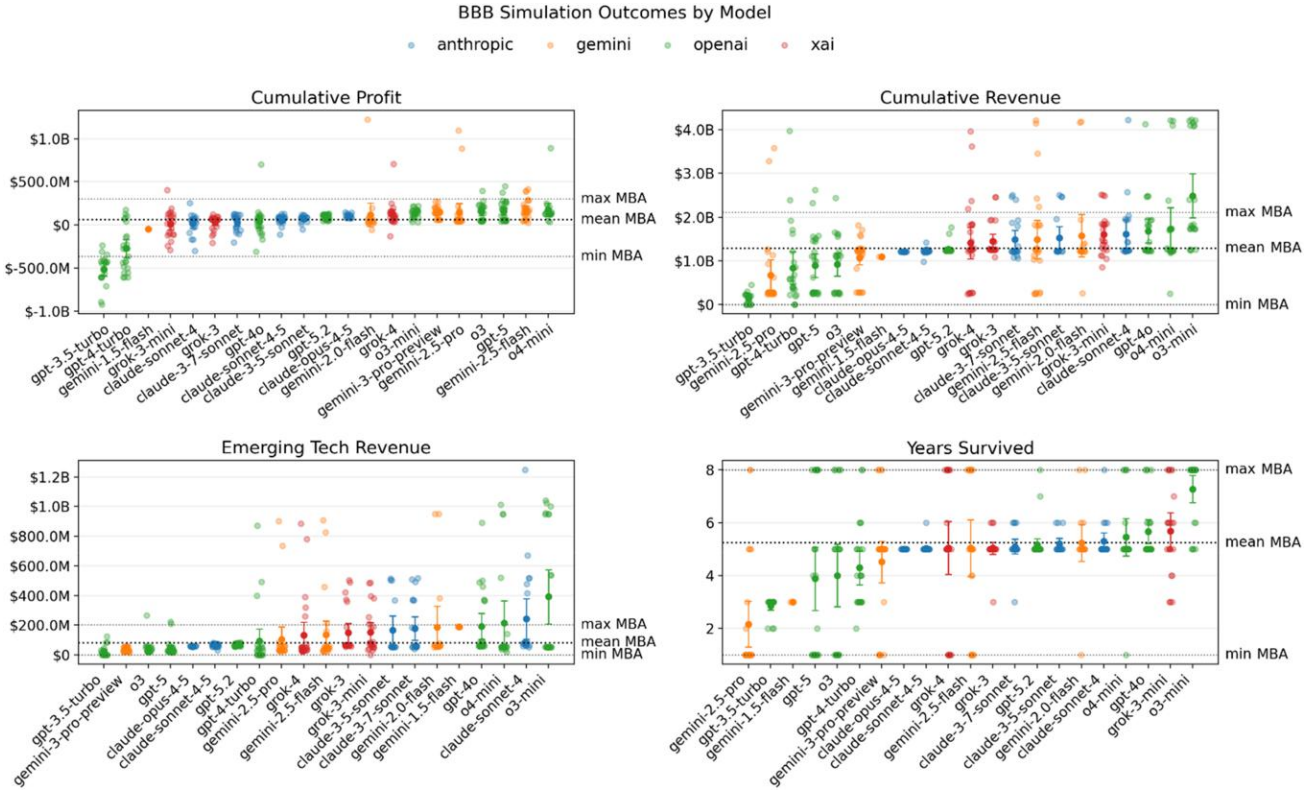
other prominent LLM benchmarks. Figure 5 plots the BBB composite score on the y axis, with a commonly reported scientific reasoning benchmark (GPQA Diamond accuracy) on the x axis. Although the more advanced models have consistently improved on the GPQA benchmark over time, Figure 5 shows a reversal in trend for the BBB composite score. From GPT-3.5 through o4-mini, the BBB score is positively correlated with improvements on GPQA diamond, but for the latest models (o3-mini through GPT-5), the trend reverses to a negative correlation. Figure 6 shows a similar story for the LM Arena benchmark, which tests human

Figure 3. (Color online) Back Bay Battery Simulation LLM Performance: Ranked Composite Scores



Notes. Each semitransparent point is the composite score from a single simulation run (horizontally jittered for readability). Solid points show the per-model mean; error bars indicate an approximate 95% confidence interval for the mean across runs. Dotted horizontal lines mark the historical MBA minimum, mean, and maximum composite scores for reference. Because the MBA results were collected under different conditions, they should be treated as contextual reference points rather than a direct apples-to-apples comparison.

Figure 4. (Color online) Back Bay Battery Simulation LLM Performance: Ranked Raw Metrics



Notes. Each semitransparent point is the score from a single simulation run (horizontally jittered for readability). Solid points show the per-model mean; error bars indicate an approximate 95% confidence interval for the mean across runs. Dotted horizontal lines mark the historical MBA minimum, mean, and maximum composite scores for reference. Because the MBA results were collected under different conditions, they should be treated as contextual reference points rather than a direct apples-to-apples comparison.

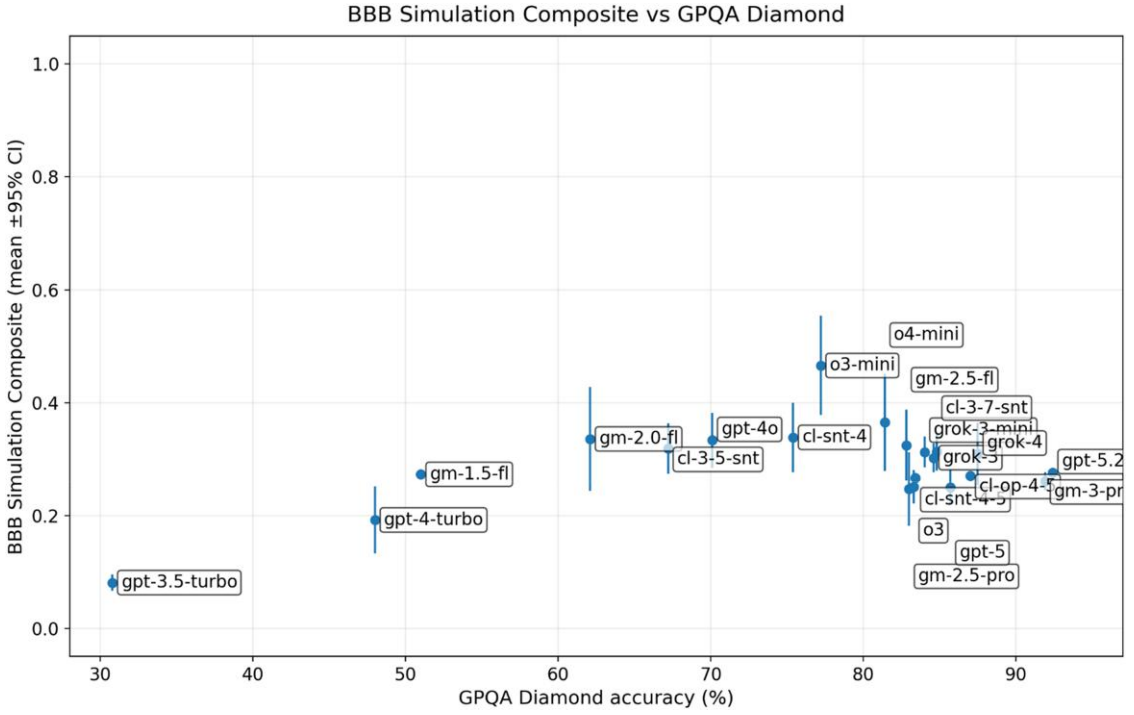
users' blind ratings of chat performance (UC Berkeley SkyLab 2025). The relationship between higher LM Arena performance and higher BBB simulation performance plateaus, or even reverses, for the latest models.

4.2. Additional Analyses: Sensitivity to Alternative Treatments and Open-Source Models

We created variations of the benchmark to explore the sensitivity of the results to alternative conditions. These results are reported in Appendix E. First, we report results from all the same models playing the basic mode of the BBB simulation. In the basic mode, it is not possible to be fired. This allowed us to test whether the surprisingly poor performance of current frontier models was due to repeated tactical errors that caused early firing or whether the same outcomes would hold under different criteria. Although there are some differences in model-specific performance,[5] the high-level outcomes are similar to the primary results: Models like GPT-5 and o3 excelled at cumulative profit but performed poorly in *Emerging Tech Revenue*. The best-performing models were still typically lighter reasoning models from late 2024 and early 2025. This confirms that the main benchmarking results were not simply an artifact of idiosyncratic firing rules in the simulation.

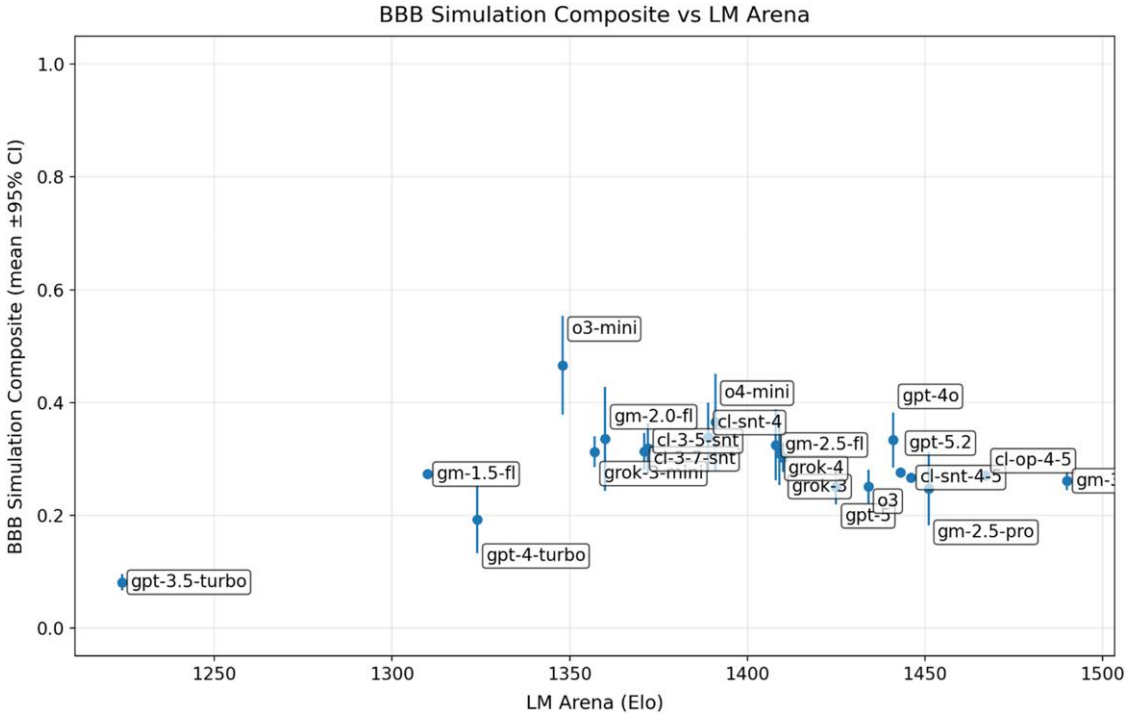
Next, we reran the primary results (in advanced mode) using an alternative prompt. Although the default prompt was outcome-agnostic ("Please make the requested strategic decisions for the upcoming year based on the information provided"), we reran the same models with additional instructions explicitly emphasizing a focus on the emerging technology.[6] This test was designed to assess whether the surprisingly poor performance of current frontier models such as GPT-5 stemmed from interpreting profit from the core business as the primary objective (Meincke et al. 2025). The results, presented in Appendix E, show that GPT-5, o3, and Gemini 2.5 Pro performed marginally better under the revised instructions than under the default prompt. However, their outcomes still lagged behind lighter models using the default prompt. Interestingly, models like o3-mini that excelled with the original open-ended instructions performed significantly worse with the emerging-technology-focused prompt. With the emerging-technology prompt, these models focused too heavily on the new technology, sacrificed

Figure 5. (Color online) BBB Simulation Composite Score vs. GPQA Diamond Benchmark



Notes. Each point indicates the per-model mean. Error bars indicate an approximate 95% confidence interval for the mean across runs. GPQA diamond scores were sourced from the GPQA leaderboard at llm-stats.com (GPQA Leaderboard 2025).

Figure 6. (Color online) BBB Simulation Composite Score vs. LM Arena Benchmark



Notes. Each point indicates the per-model mean; error bars indicate an approximate 95% confidence interval for the mean across runs. GPQA diamond scores were sourced from the LM Arena leaderboard at lmarena.ai (Overview Leaderboard|LMArena n.d.) on January 5, 2026.

core profitability, and were "fired" much earlier due to negative returns. Consequently, with the more specific emerging-technology instructions, performance exhibited a roughly linear positive correlation with model release date—but the overall average was significantly lower than with the original, less specific default prompt.

These findings suggest that the current frontier models may be more rigid and "literal" in following explicit instructions, which can become a liability in strategic or uncertain environments. When asked to optimize narrowly around emerging technology, they improved slightly—showing they were capable of shifting their focus—but their performance still fell short of lighter models under open-ended prompts. By contrast, earlier models appeared more flexible and adaptive when left with outcome-agnostic instructions, balancing short-term profitability with long-term investment. However, they were more sensitive to the prompt: when given a tightly specified goal, they overcommitted to that goal and performed on par or worse than the current frontier models.

Finally, to compare proprietary models with open-source alternatives, we also ran the simulation using prominent open-source models, including DeepSeek, Gemma, Qwen, Llama (from Meta), and GPT-OSS. Each run was conducted under the same conditions as the main analyses, with masked terms, default instructions, and Advanced mode. Results are reported in Appendix F. As shown in the figures, open-source models generally exhibited substantially lower performance on the BBB benchmark compared with proprietary systems. Their performance appears less systematically correlated with release date, and although we do not observe the clear decline seen in the latest proprietary models, open-source results seem to have plateaued at a markedly lower level of performance.

5. Discussion

This paper has presented a benchmark for the strategic decision-making capabilities of LLMs by evaluating their performance on an established business strategy simulation. Although prior research has shown that LLMs can benefit some strategy tasks (Dell'Acqua et al. 2023, 2025; Csaszar et al. 2024; Doshi et al. 2025), this is the first standardized assessment comparing multiple models' ability to independently make complex, uncertain, multiperiod business strategy decisions in a newly encountered uncertain environment (Table 1). Specifically, we benchmarked 21 proprietary and 13 open-source models from leading LLM providers on the BBB strategy simulation, which tests users' ability to synthesize complex information across multiple periods while balancing strategic investments in maintaining a profitable core business (exploitation) and in developing an emerging business (exploration).

Our results indicate that generally LLMs have significantly improved in their performance on the BBB simulation, but that the latest frontier models have declined. Early models from 2022 to 2023, such as GPT-3.5 and GPT-4, performed significantly below later models and well below the human MBA benchmark. Models from late 2024 to early 2025 (such as GPT-4o and Gemini 2.0-Flash) performed best overall, achieving strong profitability in the core market while also investing in the emerging one. These models exceeded the historical human MBA student average scores. In contrast, the current frontier models from mid-2025, such as Gemini 2.5-Pro and GPT-5, despite performing well on other benchmarks, regressed below earlier models and average historical human MBA performance.

By developing a transparent and replicable benchmarking framework, we provide a tool for strategy researchers to systematically track LLMs' evolving strategic decision-making capabilities over time.[7] Beyond establishing an evaluation of current capabilities, the benchmark allows for ongoing assessment and offers a target to improve future AI systems by highlighting the blind spots of the current frontier models. To this end, the BBB strategy benchmark is a first step—a proof of concept and a call to action for the field. Below, we distill design principles and practical recommendations for developing additional strategy-specific benchmarks. Together, these contributions are intended to guide future strategy benchmarking research, enabling scholars and developers to build, refine, and compare AI systems in ways that better reflect the distinctive demands of real-world strategic decision making.

5.1. Contributions to Research on AI and Strategy

Our study makes several contributions to research on AI in strategic decision making. First, we provide one of the first systematic, empirical benchmarks comparing multiple LLMs' independent performance in an environment that mimics real-world strategy. Prior work has tended to evaluate important albeit narrow strategy tasks such as idea generation or evaluating early business plans (Dell'Acqua et al. 2023, 2025; Csaszar et al. 2024; Doshi et al. 2025). Through the BBB simulation, we were able to interrogate a broader set of strategic challenges that mirror real-world strategy including uncertainty, complexity, irreversible multiperiod moves, and delayed/noisy feedback. Simulations like BBB, we propose, offer a rigorous, controlled, and repeatable environment that complements and extends existing approaches for evaluating LLMs.

This benchmark allowed us to demonstrate that LLMs have recently crossed a key threshold in that many models can now independently reason through complex, multiperiod, uncertain strategy problems

without supervised fine-tuning. Unlike earlier AI strategy achievements that depended on reinforcement learning across millions of data points and rich prior examples (as in chess, Go, or StarCraft) (Newborn 2012, Silver et al. 2016, Kleinberg et al. 2018, Vinyals et al. 2019, Choudhury et al. 2021, Kolkman et al. 2024), we show that frontier general-purpose LLMs can now apply strategic reasoning to unfamiliar open-ended tasks.

Such a significant expansion of AI's cognitive frontier further challenges prior scholarly assumptions about which decision-making domains remain exclusively human (Felin and Holweg 2024). Going forward, debates over AI's strategic capabilities must recognize that some models can now match the performance of MBA students in simulations that capture many widely recognized elements of "strategic" decision making. Although early models such as GPT-3.5 and GPT-4 lacked this capacity, our results suggest that later models may exceed prior expectations in managing complexity and making valuable forward-looking decisions in novel situations (Felin and Zenger 2017, Felin et al. 2024).

At the same time, the benchmark results urge caution. The current frontier models (GPT-5, Gemini 2.5 Pro, o3, etc.) were so strongly inclined to invest in the core business at the expense of emerging technologies that they performed worse than the average MBA student and only marginally better than less advanced models such as GPT-4-turbo. That their performance in the BBB simulation did not track with gains on other high-profile benchmarks like GPQA underscores the need for a strategy-specific benchmark. This points to a potential failure mode of LLMs: As they are trained and tuned for peak performance in domains such as coding, chat, or scientific reasoning, they may inadvertently converge on an inability to handle uncertainty, excelling only at replicating past data (Felin and Holweg 2024). The divergence also suggests that the microfoundations of successful strategy-making depend on cognitive abilities distinct from those required for reasoning and decision making in other domains (Felin et al. 2012, Baer et al. 2013).

Finally, another contribution is to offer a dynamic benchmarking framework that can be regularly updated over time. The fact that the BBB benchmark breaks the trend of other-domain benchmarks implies that advances in general reasoning and conversational fluency, although once correlated with better strategic performance, no longer reliably predict it. This disconnect reinforces the importance of a strategy-specific benchmark: Without it, progress in adjacent domains could be mistaken for progress in strategic decision making, leaving both scholars and practitioners with a distorted view of AI's capabilities in strategy. Furthermore, absent a strategy-specific framework, existing research on LLMs' strategic abilities risks becoming fragmented as the findings from carefully validated empirical studies become obsolete (and thus less academically or managerially useful). By providing a benchmarking infrastructure, we enable future researchers to recalibrate their assessments as AI evolves. For instance, the benchmark can track whether the decreased performance of current frontier models is a continued trend or if it is corrected over time. This benchmark and follow-ons can also influence future model developments to address this blind spot, promoting systems that are more robust for real-world strategic decision making.

5.2. Managerial Implications

Our findings have immediate implications for managers considering whether and how to integrate AI into strategic decision making (Clough and Wu 2022). Frontier AI systems have only recently shown the ability to handle complex strategic tasks once thought too ambiguous or intricate for automation without human input or extensive training. Whereas this was not true even in early 2024, it is now possible for some models to augment decisions involving technology or market transitions, allocation of scarce R&D budgets across competing opportunities, and the balance between exploitation and exploration. More than a tactical tool, some LLMs can increasingly be viewed as a candidate to support higher-order strategic functions, particularly in structured decision environments.

At the same time, the benchmark highlights that the current frontier models remain rigid and risk averse, consistently favoring investment in the core business while avoiding emerging technologies. The benchmark thus reveals which models perform well (such as Grok 3, GPT-4o, o3-mini, Claude Sonnet 4) and which perform surprisingly poorly (such as GPT-5 or Gemini 2.5 Pro) in strategic contexts. Although performance varies depending on the precise setup, the evidence makes clear that advancement on general benchmarks in chat, coding, or science does not necessarily translate into more effective strategic decision making. Any individual model may carry unknown risks. One promising approach is to use ensembles of different models (Doshi et al. 2025) or to triangulate among them as with other methods (Hong et al. 2021, Allen and McDonald 2025).

5.3. Guidance for Future Strategy Benchmarks

Although BBB offers many advantages for testing LLM strategic decision making, it also has notable gaps. We view BBB as a necessary first step toward building benchmarking infrastructure in our field, spurring a call to action and a guide for future work. Looking ahead, we propose a simple taxonomy of desirable characteristics for future strategy benchmarks to form the basis of a research agenda. We organize this

taxonomy using the four categories in Table 1: Elements of Strategy, Business Context, Interaction Structure, and AI Characteristics.

5.3.1. Elements of Strategy. Future strategy benchmarks should strive to capture the core elements of strategic decision making discussed in this paper: complexity, multiperiod dynamics, irreversible commitments, delayed and noisy feedback, and uncertainty/novelty. Benchmarks that do not capture these elements (such as a benchmark that is merely a Q&A of strategy concepts) will not add much value beyond existing benchmarks in other domains (such as the GPQA Diamond PhD-level science benchmark). BBB does incorporate these strategy elements, but in a single setting and with a relatively fixed "black box" structure. In the future, additional strategy simulation benchmarks could vary parameters like the number and interdependence of strategic levers or the length and structure of time horizons. Benchmarks could also intentionally build in more genuine novelty—problems that cannot be solved by analogy to familiar cases (Fleming et al. 2007, McDonald and Allen 2022, Morris et al. 2023, Allen 2025). Designing families of scenarios that span these dimensions would make it possible to test whether a model's "strategy capability" generalizes beyond any one game or case.

5.3.2. Business Context. Because LLM behavior is highly context specific, these benchmarks should be naturalistic—rooted in realistic narratives, data, and tradeoffs that resemble actual managerial problems—while remaining sufficiently structured to allow transparent scoring and comparison. BBB focuses on a specific context: strategic resource allocation to new technology in an established firm. Future benchmarks can extend to other business contexts and stages, including alternative industry structures, early-stage innovation, scaling and platform competition, turnaround situations, and settings where the primary outcome is not purely financial. For example, future benchmarks could consider how LLMs prioritize and navigate other outcomes such as ethical considerations, external stakeholder buy-in, regulatory constraints, or mission objectives alongside profitability and growth.

5.3.3. Interaction Structure. A central dimension for future strategy benchmarks is multiagent dynamics. In BBB, competition is implicit in the market response built into the simulation, and the focal decision maker does not interact directly with rivals, partners, or stakeholders. Future benchmarks could test richer multiplayer simulations, including model-versus-model and model-versus-human play, where performance depends on anticipating and influencing other agents' behavior. Such environments would allow researchers to study escalation, deterrence, cooperation, and collusion, as well as path-dependent outcomes that emerge from repeated interaction (for examples of simulations with such dynamics, see Chen et al. (2010) and Eisenmann (2025)). Multiplayer simulations might even unfold in an "arena" where LLMs compete against one another (UC Berkeley SkyLab 2025), generating Elo-style scores for strategy-specific games and contexts.

In addition, future benchmarks could also probe agentic abilities: models playing in the same interface as humans, deciding what information to seek, when to act, and how to manage situations in real time for a sustained period of time (in BBB, this information was fed to LLMs by a structured program). Today's LLMs mostly lack these long-term agentic capabilities, but progress along this dimension would be a major milestone (Kwa et al. 2025).

5.3.4. AI Characteristics. Finally, future benchmarks should be sensitive to different AI characteristics, including architecture and training regime. We emphasize that reinforcement learning approaches used in games like AlphaStar and chess (playing millions of structured games as training data) are not useful for assessing real-world strategy capabilities. However, future benchmarks could explore how LLMs (or later AI systems) perform when fine-tuned on other strategy cases (e.g., a corpus of strategy cases and teaching notes) or through interactive learning from human feedback. The BBB benchmark only tests general out-of-the-box LLM capabilities, but focusing attention through fine-tuning or prompting could yield very different results.

5.3.5. Research Agenda. Taken together, we suggest a future research agenda centered on creating open-source strategy simulations that give researchers fine-grained control over benchmark design and evaluation. Rather than relying on proprietary or opaque environments, we suggest that modular, extensible simulation frameworks could allow systematic variation in strategic complexity, temporal dynamics, interaction structures, and outcome measures to isolate specific capabilities and test generalization across contexts. A key challenge is to construct benchmark environments that remain replicable and contamination resistant at scale, for example, through procedurally generated scenarios with randomized but balanced parameters, application-based access to the open-source simulations (to avoid leaking to LLMs), periodic refreshes of market structures, and masking of surface details so that solutions cannot simply be memorized. This agenda also calls for close collaboration between AI researchers and strategy scholars to ensure that benchmarks capture the conceptual richness of strategic reasoning rather than merely testing pattern

matching or statistical optimization. By advancing such a shared experimental infrastructure, the field can move toward more rigorous, transparent, and cumulatively informative evaluations of strategic capabilities in AI.

5.4. Other Limitations and Future Research

Because AI remains far from functioning as an independent strategist, future research should also examine how humans use AI within simulations. Current systems lack physicality, social presence, agency, and long-term task management, making them ill suited for real-world strategic automation (Murray et al. 2021, Randazzo et al. 2025). In practice, applications will take the form of human–AI collaboration rather than full substitution (Mollick 2024, Randazzo et al. 2025). This raises a critical question: Would a current frontier model such as GPT-5 perform more effectively when guided by a skilled strategist providing well-framed prompts, or would it simply replicate the same failure modes it displays when used alone? Because most strategic work is team based, another key question is how group dynamics shape these interactions. How would teams of humans—or ensembles of LLMs—perform under different conditions? Such comparisons could reveal whether the systematic biases we observe (e.g., the exploitation tendency of frontier models) are mitigated or amplified in group settings. Future research should also compare AI and human performance under more controlled conditions, as the human data in this study derive from historical student performance rather than tightly controlled laboratory experiments. Variations in prompting, task setup, and framing may all influence outcomes for both humans and AI. We hope this study provides a foundation from which to launch such future investigations.

Appendix A. Human Performance Benchmarks Each Year

Year	*n*	Statistic	Cumulative profit ($M)	Cumulative revenue ($M)
Overall	249	Mean	62.1	1,335.0
		Standard error	5.1	27.0
2018	81	Mean	68.7	1,425.8
		Standard error	8.1	49.6
2020	88	Mean	57.1	1,371.7
		Standard error	9.2	39.5
2023	19	Mean	62.5	1,158.8
		Standard error	16.8	103.8
2024	20	Mean	51.1	1,160.3
		Standard error	17.8	114.2
2025	41	Mean	65.0	1,243.9
		Standard error	12.9	62.0

Note. The earlier years included multiple large class sections, whereas the later years included only one or two small sections.

Appendix B. AI-Simulation Interface Pseudocode

```
FOR each (model, prompt_style, difficulty) IN experiment_matrix:
    reset_simulation()
    FOR run = 1 TO N_runs:
        driver ← login_to_HBSP()
        FOR year = 3 TO 10:

            ### 1. COLLECT DATA ###
            switch_to("Analyze")
            wait_until_loading_finishes()
            data   ← scrape_subtabs() + scrape_news()
            budget ← detect_RD_ceiling()

            ### 2. BUILD PROMPT ###
            prompt ← [BACKGROUND, FORMAT_RULES,
                      INSTRUCTION_TEMPLATE(year),
                      DATA_BLOCK(data, budget),
                      PAST_DECISIONS(run, year - 1)]

            ### 3. CALL LLM ###
            reply     ← openai.chat_completion(model, prompt)
            decisions ← parse_numbers(reply)
            IF invalid_format OR ΣR&D > budget:
                retry ≤ 3 times with corrective messages
            IF still invalid:
                abort run

            ### 4. SUBMIT DECISIONS ###
            type_prices_units_R&D(decisions)
            click_submit()
            IF game_over():
                BREAK loop_years

        ### 5. END-OF-RUN HOUSEKEEPING ###
        record_best_scores()
        save_run_results_to_JSON()
        reset_simulation()

consolidate_logs_and_results()
```

Appendix C. LLM Model Aliases and Dates Run

Model (name used)	Run date	Simulation difficulty mode
openai/gpt-3.5-turbo-0125	8/1/2025	Advanced
openai/gpt-4-turbo-2024-04-09	8/1/2025	Advanced
openai/gpt-4o-2024-08-06	8/1/2025	Advanced
anthropic/claude-3-5-haiku-20240307	8/2/2025	Advanced
anthropic/claude-3-5-sonnet-20240620	8/2/2025	Advanced
anthropic/claude-3-7-sonnet-20250219	8/2/2025	Advanced
anthropic/claude-sonnet-4-20250514	8/2/2025	Advanced
openai/o3-2025-04-16	8/2/2025	Advanced
openai/o3-mini-2025-01-31	8/2/2025	Advanced
openai/o4-mini-2025-04-16	8/2/2025	Advanced
xai/grok-3-mini	8/2/2025	Advanced
gemini/gemini-1.5-flash	8/3/2025	Advanced
gemini/gemini-2.0-flash	8/3/2025	Advanced
gemini/gemini-2.5-flash	8/3/2025	Advanced
xai/grok-3	8/3/2025	Advanced
xai/grok-4-0709	8/3/2025	Advanced
openai/gpt-5-2025-08-07	8/12/2025	Advanced
gemini/gemini-2.5-pro	8/14/2025	Advanced
openai/gpt-3.5-turbo-0125	8/15/2025	Basic
openai/gpt-4-turbo-2024-04-09	8/15/2025	Basic
openai/gpt-4o-2024-08-06	8/16/2025	Basic
openai/gpt-5-2025-08-07	8/16/2025	Basic
openai/o3-2025-04-16	8/16/2025	Basic
openai/o3-mini-2025-01-31	8/16/2025	Basic
openai/o4-mini-2025-04-16	8/16/2025	Basic
anthropic/claude-3-5-sonnet-20240620	8/17/2025	Basic
anthropic/claude-3-7-sonnet-20250219	8/17/2025	Basic
anthropic/claude-sonnet-4-20250514	8/17/2025	Basic
xai/grok-3	8/17/2025	Basic
xai/grok-3-mini	8/17/2025	Basic
xai/grok-4-0709	8/17/2025	Basic

Appendix C. (Continued)

Model (name used)	Run date	Simulation difficulty mode
gemini/gemini-1.5-flash	8/18/2025	Basic
openai/gpt-5.2-2025-12-11	12/17/2025	Advanced
anthropic/claude-sonnet-4-5-20250929	12/17/2025	Advanced
anthropic/claude-opus-4-5-20251101	12/17/2025	Advanced
gemini/gemini-3-propreview	12/17/2025	Advanced
openai/gpt-5.2-2025-12-11	12/30/2025	Basic
anthropic/claude-sonnet-4-5-20250929	12/30/2025	Basic
anthropic/claude-opus-4-5-20251101	12/30/2025	Basic
gemini/gemini-3-propreview	12/30/2025	Basic

Appendix D. Masking Procedure to Mitigate Contamination

Case-insensitive replacement terms used in "masking" function:

- **Simulation/company name**
  - "Back Bay Battery" → **EnergyCo**
  - "Back Bay" → **EnergyCo**
- **Technologies & acronyms**
  - "Absorbed Glass Mat" → **AeroBond Matrix**
  - "AGM" → **ABM**
  - "Supercapacitor" → **Quantum Storage Cell**
  - "Supercapacitors" → **Quantum Storage Cells**
  - "SC" → **QSC**
- **Industry segments**
  - "Warehouse equipment" → **Industrial Handling Systems**
  - "IHS" → **IHS** (normalized/capitalized)
  - "Uninterruptible power system(s)/supply(ies)" (optionally "(UPS)") → **Continuity Power Modules**
  - "UPS" → **CPM**
  - "UPSs" → **CPMs**
  - "automobile" → **Light Transport Vehicle**
  - "automobiles" → **Light Transport Vehicles**
  - "automotive" → **LTV sector**
- **Performance dimensions (in prose/tables)**
  - "Energy Density" → **Specific Energy Index (SEI)**
  - "Recharge Cycle(s)" → **Cycle Endurance Rating (CER)**
  - "Self-discharge" → **Standby Loss Interval (SLI)**
  - "Recharge Time" → **Recovery Time Index (RTI)**
  - "Recharge Duration" → **Recovery Time Index (RTI)**

Figure D.1. Masked Background Briefing

You are the head of EnergyCo, Inc., a division generating $240 million in annual sales as part of a larger $40 billion consumer electronics conglomerate. EnergyCo specializes in manufacturing two battery variants: the AeroBond Matrix (ABM) battery and the recently introduced Quantum Storage Cell (QSC) battery. ABM batteries are robust, sealed lead-acid units designed to withstand harsh environments and vibrations, accounting for approximately 80% of the firm's revenue, serving as the cornerstone of its operations. The newer Quantum Storage Cell batteries account for the remaining 20% of sales.

EnergyCo markets its batteries exclusively to commercial clients across three key sectors:

Light Transport Vehicles (LTVs)
EnergyCo's ABM batteries have several qualities making them ideal for premium vehicles, as their sealed design allows flexible placement within vehicle interiors or trunks. The growing adoption of start-stop technology, which enhances fuel efficiency, has boosted automaker demand for ABM batteries due to their suitability for the increased electrical loads and frequent cycling. Additionally, ABM batteries are increasingly utilized in regenerative braking systems, where generators integrated with the drivetrain of popular European SUVs capture braking energy to power vehicle electronics, further improving fuel economy.

Industrial Handling Systems (IHS)
EnergyCo has established a niche supplying batteries to automated logistics facilities, where equipment such as stackers and robotic systems heavily rely on electric power. These systems typically experience significant startup power demands and require regular recharge cycles, providing a growing market segment for EnergyCo's battery offerings.

Continuity Power Modules (CPM)
This sector encompasses emergency backup power solutions for large-scale data centers. EnergyCo's batteries fill warehouse-sized spaces to deliver immediate power during outages, bridging the roughly 15-second interval necessary to activate standby diesel generators.

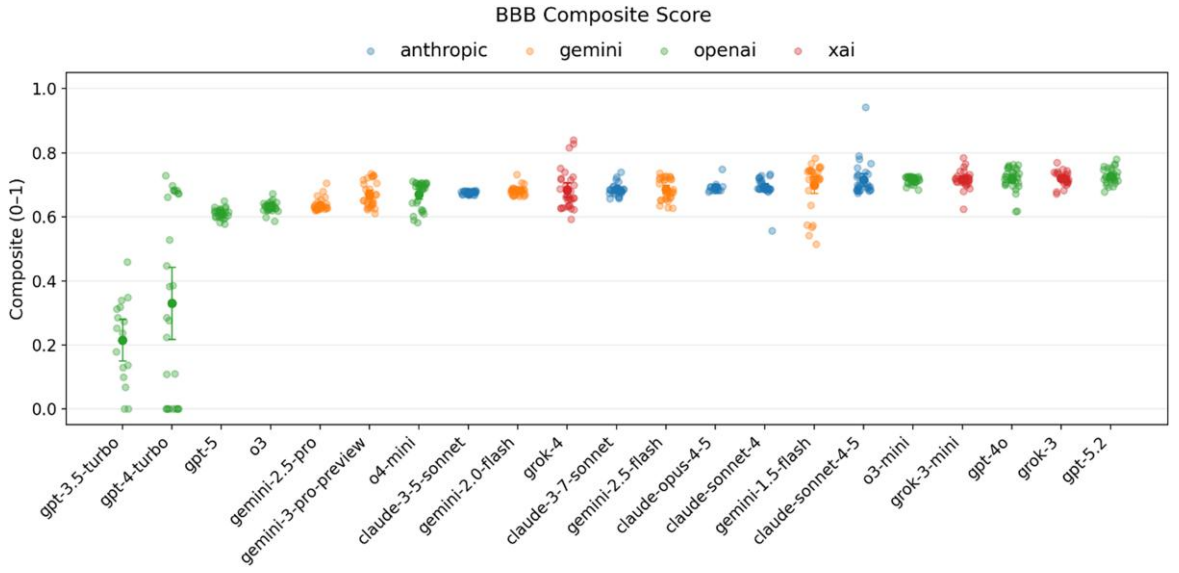
Each battery type has differing strengths across five performance metrics: Specific Energy Index, Cycle Endurance Rating, Standby Loss Interval, recharge duration, and Recovery Time Index. Market segments differ in prioritizing these features.

Continuous pricing pressures restrict EnergyCo's available R&D funds, making prioritization critical. The sales department consistently advocates for improvements in LTV battery performance, as minor technical enhancements can significantly influence order outcomes in this price-sensitive market. Price adjustments can influence demand, but the company must cautiously avoid losing major customers. Additionally, EnergyCo must carefully consider investment lead times before benefits materialize. Although some researchers are enthusiastic about Quantum Storage Cell potential, the technology currently falls short in meeting essential requirements for EnergyCo's primary markets, necessitating careful management to avoid financial risk.

The product lead at a prominent automaker—one of EnergyCo's biggest customers—has urged the company to tailor developments specifically for an upcoming SUV model update, emphasizing higher power density and lower battery costs in response to intense competition from Asian suppliers. This customer is already evaluating offshore ABM battery suppliers and has emphasized the necessity for EnergyCo to maintain competitive pricing. Although Quantum Storage Cells offer attractive quick recharge capabilities, their limited storage capacity is a significant drawback. Prioritizing this automaker's requirements would essentially monopolize EnergyCo's limited R&D resources.

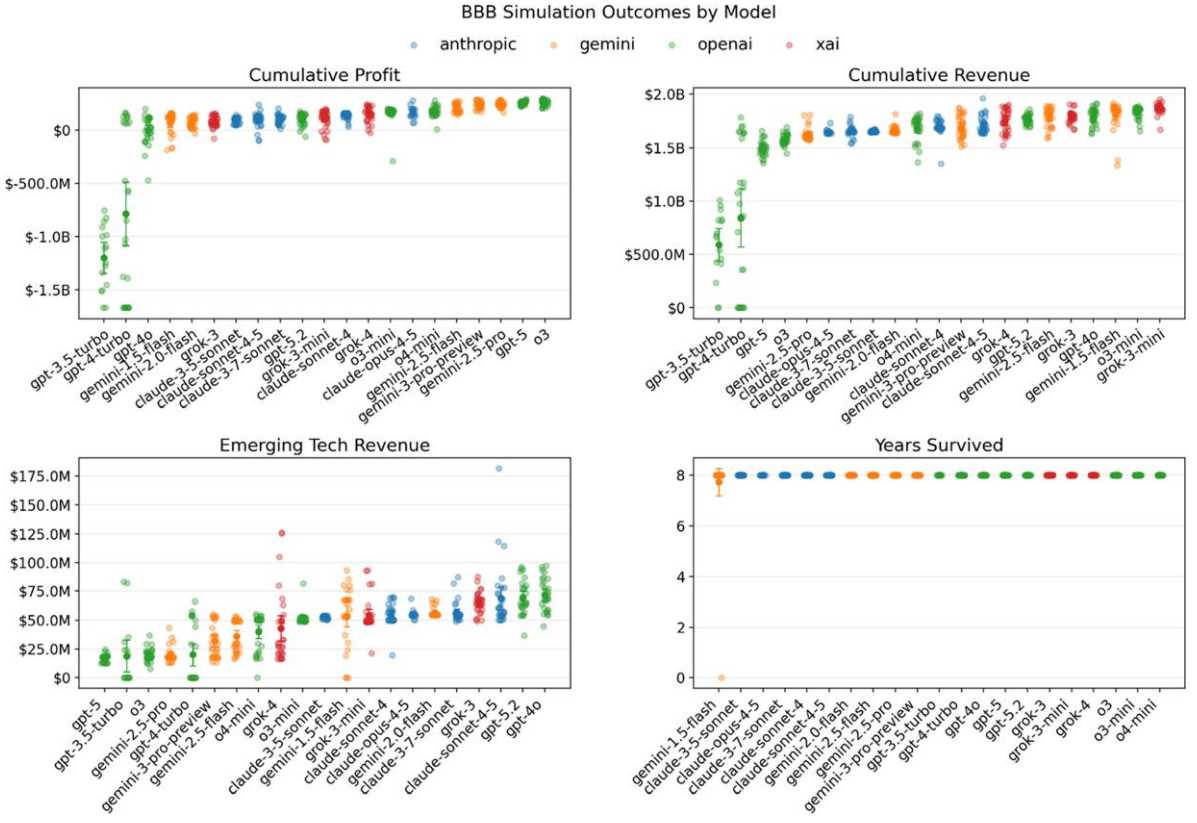
Appendix E. Alternative Specifications

Figure E.1. (Color online) "Basic" Mode (No Firing) Composite Results: Ranked by Score



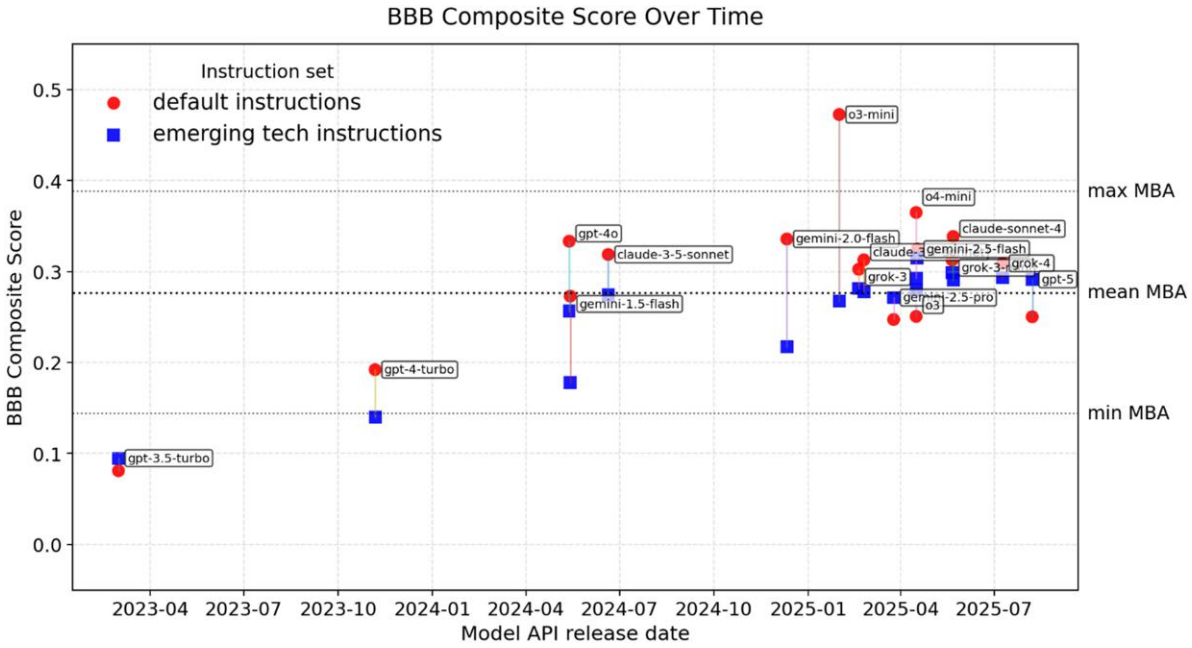
Notes. Each semitransparent point is the score from a single simulation run (horizontally jittered for readability). Solid points show the per-model mean; error bars indicate an approximate 95% confidence interval for the mean across runs.

Figure E.2. (Color online) "Basic" Mode (No Firing) Raw Metrics: Ranked by Score



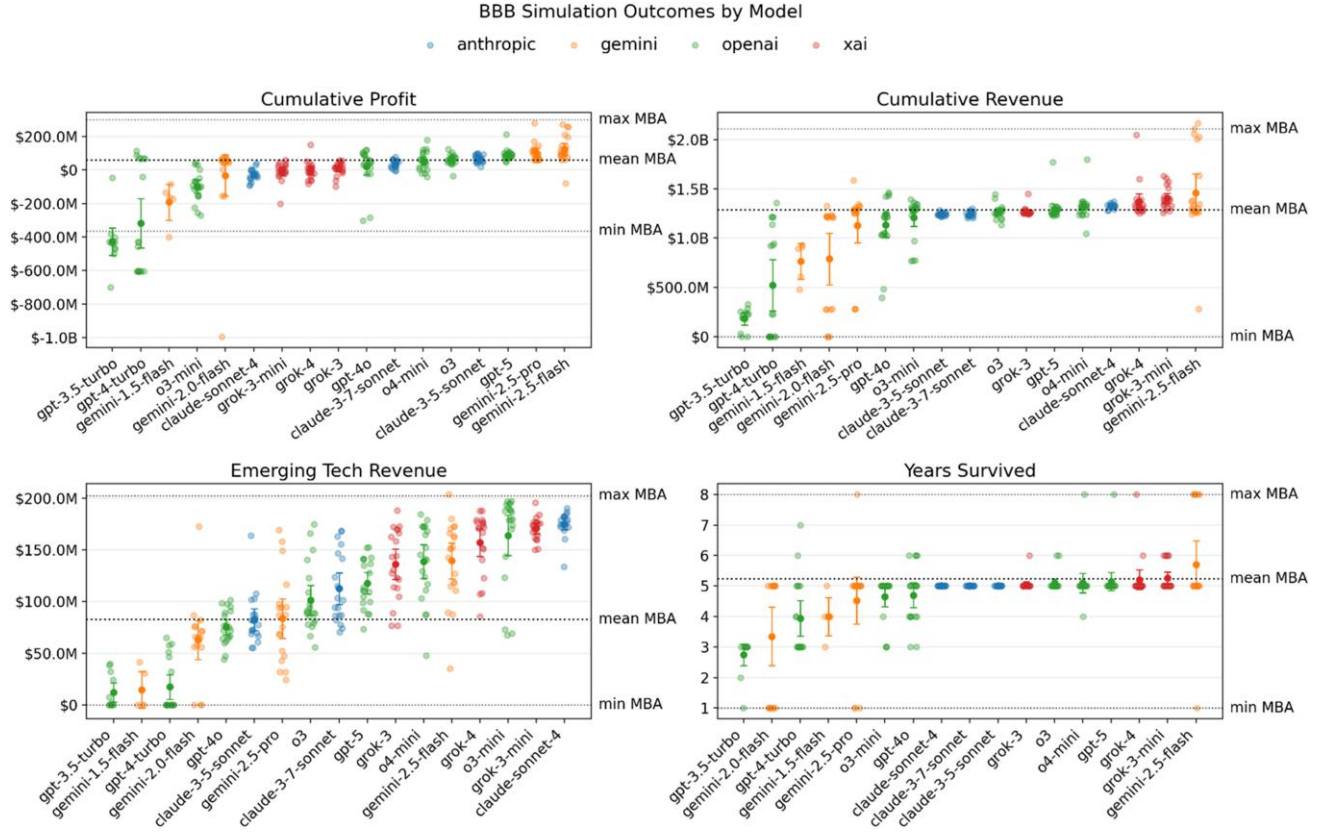
Notes. Each semitransparent point is the score from a single simulation run (horizontally jittered for readability). Solid points show the per-model mean; error bars indicate an approximate 95% confidence interval for the mean across runs.

Figure E.3. (Color online) Emerging Tech Focused Prompt: Composite Scores over Time



Notes. Each point indicates the per-model mean composite score. Dotted horizontal lines mark the historical MBA minimum, mean, and maximum composite scores for reference. Because the MBA results were collected under different conditions, they should be treated as contextual reference points rather than a direct apples-to-apples comparison. Each point indicates the per-model mean composite score, comparing the default instructions (circles) to the emerging technology instructions (squares) for each model.

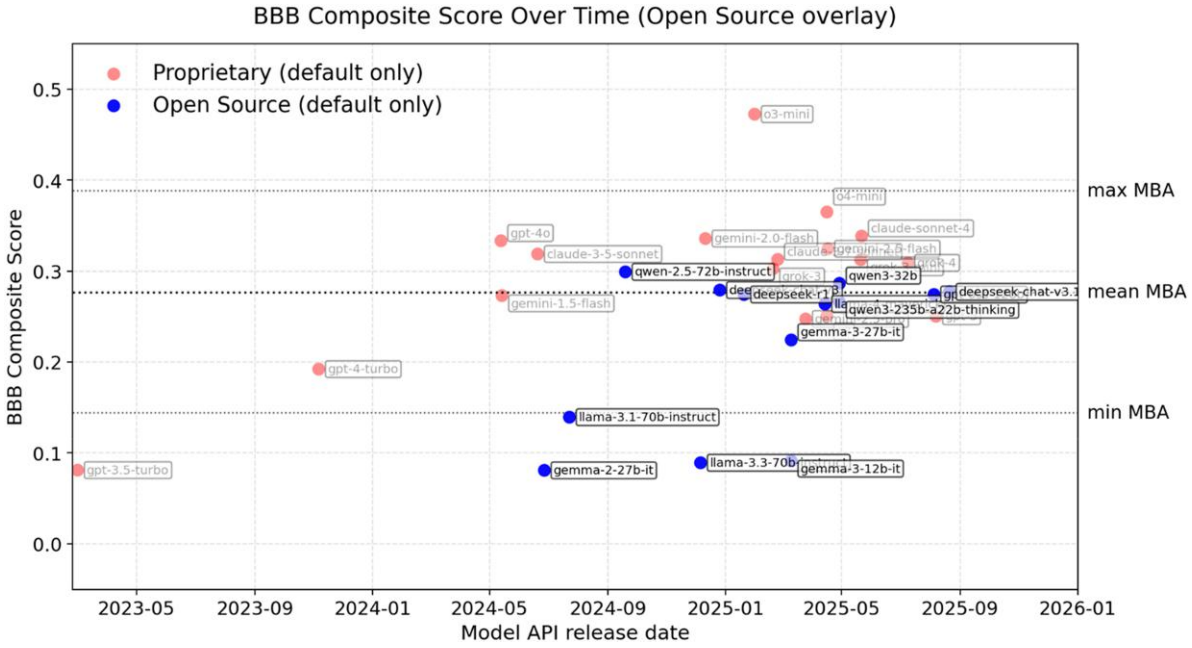
Figure E.4. (Color online) Emerging Tech Focused Prompt: Ranked Raw Metrics



Notes. Each semitransparent point is the score from a single simulation run (horizontally jittered for readability). Solid points show the per-model mean; error bars indicate an approximate 95% confidence interval for the mean across runs.

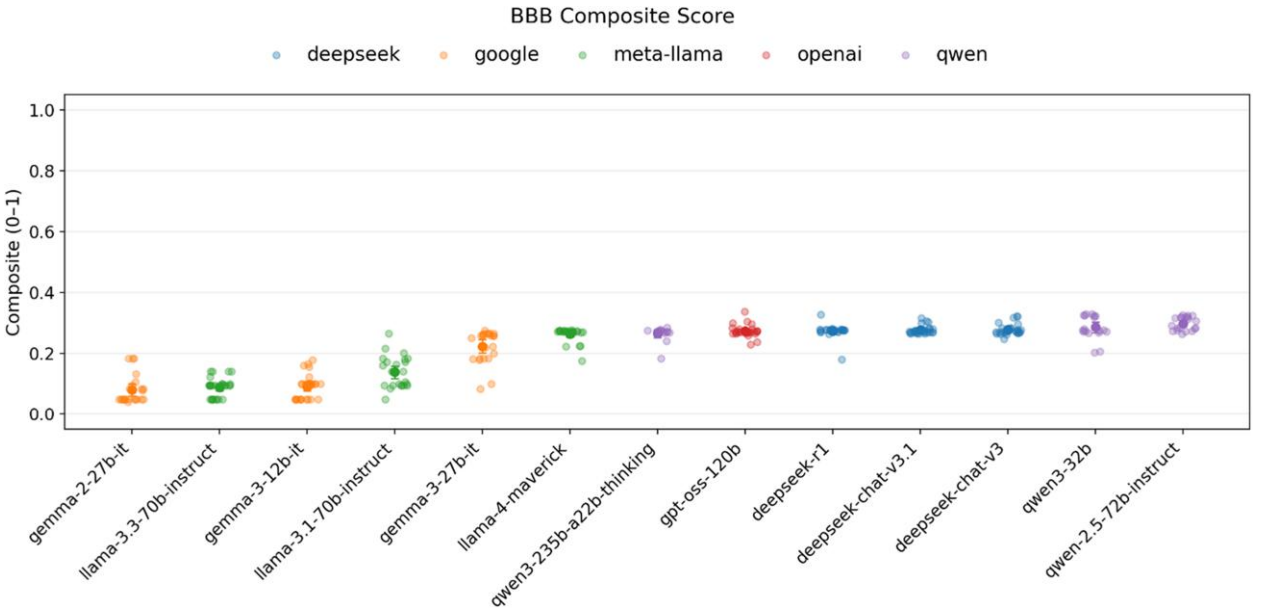
Appendix F. Open-Source Model Results

Figure F.1. (Color online) Composite Results for Open-Source Models (Masked; Advanced Mode; Default Instructions) over Time



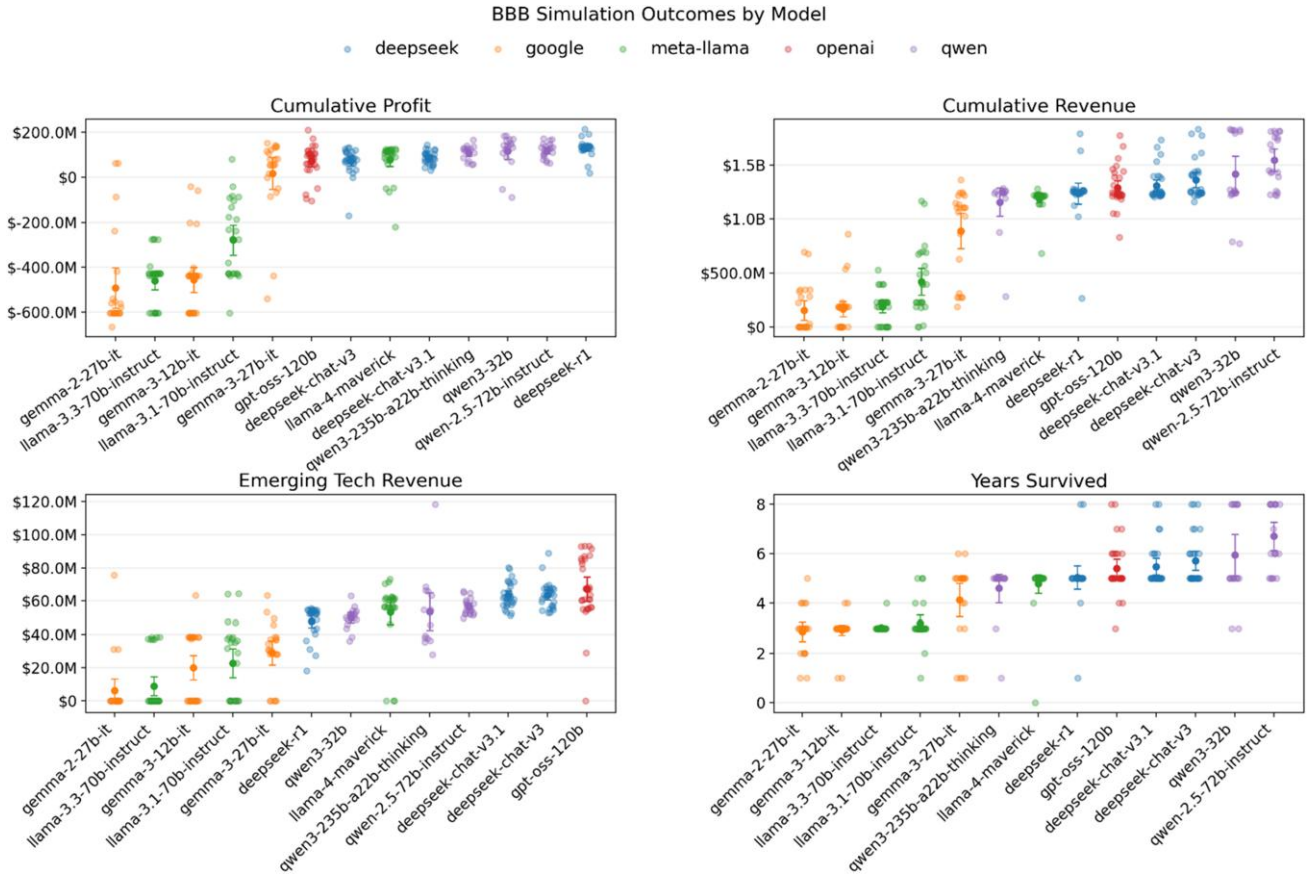
Notes. Each point indicates the per-model mean composite score, comparing the open-source models (opaque) to the proprietary models (transparent) for each model. Historical MBA scores represented as dashed lines to serve as reference points.

Figure F.2. (Color online) Composite Results for Open-Source Models (Masked; Advanced Mode; Default Instructions) Ranked by Score



Notes. Each point indicates the per-model mean composite score, comparing the open-source models (opaque) to the proprietary models (transparent) for each model. Historical MBA scores represented as dashed lines to serve as reference points.

Figure F.3. (Color online) Raw Metrics for Open-Source Models (Masked; Advanced Mode; Default Instructions) Ranked by Score



Notes. Each point indicates the per-model mean composite score, comparing the open source models (opaque) to the proprietary models (transparent) for each model. Historical MBA scores represented as dashed lines to serve as reference points.

Endnotes

[1] Like GPQA for PhD-level science questions and LM Arena for user-rated chat performance.

[2] See the latest results at https://ryantallen.github.io/bbb_benchmark_public/.

[3] This limitation applies only to the min–max normalization step: Composite scores are comparable across data sets only when they are normalized using the same reference set (i.e., the same min and max values). When computed from a shared normalization baseline, composite scores can be compared across data sets. By contrast, the underlying raw outcomes (e.g., cumulative revenue and cumulative profit) remain directly comparable across data sets, provided they are generated under the same game and difficulty mode.

[4] We used all available data and did not omit any years. Some years did not have any simulation results available for collection.

[5] Notably, there were two exceptions in this robustness check. In the latest Basic-mode run, GPT-5.2 performed substantially better and ranked first overall, and o3-mini (the top performer in Advanced) performed noticeably worse. o3-mini's Advanced-mode strength often came less from aggressive investment in the emerging SC technology and more from conservative pricing choices that reduced the probability of being fired; in many runs it still finished with ≈$50M in SC revenue despite minimal SC investment. In Basic (where firing is impossible) this "risk management" advantage disappears, and similarly aged models that invest more heavily in SC tend to outperform it. Outside of these exceptions, relative performance for the remaining models closely tracks the Advanced-mode results, including the recurring tradeoff in which frontier models perform strongly on cumulative profit but lag on *Emerging Tech Revenue*.

[6] Specifically, the emerging technology-focused prompt was "Please make the requested strategic decisions for the upcoming year based on the information provided. IMPORTANT: your goal is to make investments to maximize revenue from the emerging technology by the end of year 10, while still maintaining profitability in existing markets."

[7] The code repository to replicate these results is available at https://github.com/ryantallen/bbb_benchmark_public.

References

Adner R (2002) When are technologies disruptive? A demand-based view of the emergence of competition. *Strategic Management J.* 23:667–688.

Adner R, Zemsky P (2005) Disruptive technologies and the emergence of competition. *RAND J. Econom.* 36(2):229–254.

Adner R, Csaszar FA, Zemsky PB (2014) Positioning on a multiattribute landscape. *Management Sci.* 60(11):2794–2815.

Allen RT (2025) Leap of faith? How diffusion dynamics obfuscate the commercial potential of novel innovations. Preprint, submitted February 14, http://dx.doi.org/10.2139/ssrn.5084612.

Allen RT, Choudhury P. (2022) Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organ. Sci.* 33(1):149–169.

Allen RT, McDonald RM (2025) Methodological pluralism and innovation in data-driven organizations. *Admin. Sci. Quart.* 70(2): 403–443.

Allen RT, Bremner R, McDonald RM (2026) Listen to your users? Self-selection in user community feedback and commercial success. *Acad. Management J.* Forthcoming.

Anthropic (2024) Introducing the next generation of Claude. https://www.anthropic.com/news/claude-3-family.

Arthur F, Hossein KR (2019) Deep learning in medical image analysis: A third eye for doctors. *J. Stomatology Oral Maxillofacial Surgery* 120(4):279–288.

Baer M, Dirks KT, Nickerson JA (2013) Microfoundations of strategic problem formulation. *Strategic Management J.* 34(2):197–214.

Benner MJ, Tushman ML (2003) Exploitation, exploration, and process management: The productivity dilemma revisited. *Acad. Management Rev.* 28(2):238–256.

Boussioux L, Lane JN, Zhang M, Jacimovic V, Lakhani KR (2024) The crowdless future? Generative AI and creative problem-solving. *Organ. Sci.* 35(5):1589–1607.

Bubeck S, Chadrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, et al. (2023) Sparks of artificial general intelligence: Early experiments with GPT-4. Preprint, submitted March 22, https://arxiv.org/abs/2303.12712.

Chen EL, Katila R, McDonald R, Eisenhardt KM (2010) Life in the fast lane: Origins of competitive interaction in new vs. Established markets. *Strategic Management J.* 31(13):1527–1547.

Chen M, Tworek J, Jun H, Yuan Q, Pinto HPDO, Kaplan J, Edwards H, et al. (2021) Evaluating large language models trained on code. Preprint, submitted July 7, https://arxiv.org/abs/2107.03374.

Choudhury P, Allen RT, Endres MG (2021) Machine learning for pattern discovery in management research. *Strategic Management J.* 42(1):30–57.

Choudhury P, Starr E, Agarwal R (2020) Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management J.*

Christensen CM, Bower JL (1996) Customer power, strategic investment, and the failure of leading firms. *Strategic Management J.* 17:197–218.

Christensen CM, Shih WC (2019) *Strategic Innovation Simulation: Back Bay Battery v3* (Harvard Business School Publishing, Cambridge, MA).

Christensen CM, McDonald R, Altman EJ, Palmer JE (2018) Disruptive innovation: An intellectual history and directions for future research. *J. Management Stud.* 55(7):1043–1078.

Chu LY, Li G, Wu A, Wu B (2025) Disruptive timing. *Management Sci.*, ePub ahead of print October 30, https://doi.org/10.1287/mnsc.2023.01734.

ClaudePlaysPokemon (2025) Retrieved August 18, 2025, https://www.twitch.tv/claudeplayspokemon.

Clough DR, Wu A (2022) Artificial intelligence, data-driven learning, and the decentralized structure of platform ecosystems. *Acad. Management Rev.* 47:184–189.

Csaszar FA (2018) What makes a decision strategic? *Strategy Sci.* 3:606–619.

Csaszar FA (2025) Unbounding rationality: Why AI is a fundamental issue for strategy. Preprint, submitted September 8, https://doi.org/10.2139/ssrn.5454634.

Csaszar FA, Levinthal DA (2016) Mental representation and the discovery of new strategies. *Strategic Management J.* 37:2031–2049.

Csaszar FA, Ketkar H, Kim H (2024) Artificial intelligence and strategic decision-making: Evidence from entrepreneurs and investors. *Strategy Sci.* 9:322–345.

Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, Lakhani KR (2023) Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Preprint, submitted September 18, https://doi.org/10.2139/ssrn.4573321.

Dell'Acqua F, Ayoubi C, Lifshitz-Assaf H, Sadun R, Mollick ER, Mollick L, Han Y, Goldman J, Nair H, Taub S, Lakhani KR (2025) The cybernetic teammate: A field experiment on generative AI reshaping teamwork and expertise. NBER Working Paper No. 33641, National Bureau of Economic Research, Cambridge, MA.

Doshi AR, Bell JJ, Mirzayev E, Vanneste BS (2025) Generative artificial intelligence and evaluating strategic decisions. *Strategic Management J.* 46:583–610.

Eisenhardt KM (1989) Making fast strategic decisions in high-velocity environments. *Acad. Management J.* 32:543–576.

Eisenhardt KM, Bourgeois LJ III (1988) Politics of strategic decision making in high-velocity environments: Toward a midrange theory. *Acad. Management J.* 31:737–770.

Eisenhardt KM, Zbaracki MJ (1992) Strategic decision making. *Strategic Management J.* 13:17–37.

Eisenmann T (2025) *Scaling Tech Ventures Simulation* (Harvard Business School Publishing, Cambridge, MA).

Felin T, Holweg M (2024) Theory is all you need: AI, human cognition, and causal reasoning. *Strategy Sci.* 9:346–371.

Felin T, Zenger T (2017) The theory-based view: Economic actors as theorists. *Strategy Sci.* 2:258–271.

Felin T, Gambardella A, Zenger T (2024) Theory-based decisions: Foundations and introduction. *Strategy Sci.* 9:297–310.

Felin T, Foss NJ, Heimeriks KH, Madsen TL (2012) Microfoundations of routines and capabilities: Individuals, processes, and structure. *J. Management Stud.* 49:1351–1374.

Fleming L, Mingo S, Chen D (2007) Collaborative brokerage, generative creativity, and creative success. *Admin. Sci. Quart.* 52: 443–475.

Gaessler F, Piezunka H (2023) Training with AI: Evidence from chess computers. *Strategic Management J.* 44:2724–2750.

Ghemawat P (1991) *Commitment: The Dynamics of Strategy* (Free Press, New York).

Google (2025) Gemini 2.5: Our most intelligent AI model. (March 25), https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-model-thinking-updates-march-2025/.

GPQA Leaderboard (2025) Retrieved December 17, 2025, https://llm-stats.com/benchmarks/gpqa.

Gupta AK, Smith KG, Shalley CE (2006) The interplay between exploration and exploitation. *Acad. Management J.* 49:693–706.

Hong L, Lamberson PJ, Page SE (2021) Hybrid predictive ensembles: Synergies between human and computational forecasts. *J. Soc. Comput.* 2:89–102.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. *Quart. J. Econom.* 133:237–293.

Kolkman D, Lee GK, van Witteloostuijn A (2024) Data science and automation in the process of theorizing: Machine learning's power of induction in the co-duction cycle. *PLoS One* 19(11):e0309318.

Kwa T, West B, Becker J, Deng A, Garcia K, Hasin M, Jawhar S (2025) Measuring AI ability to complete long tasks. Preprint, submitted March 18, https://arxiv.org/abs/2503.14499.

Levinthal DA (1997) Adaptation on rugged landscapes. *Management Sci.* 43:934–950.

Levinthal DA (2017) Mendel in the c-suite: Design and the evolution of strategies. *Strategy Sci.* 2:282–287.

Li D (2017) Expertise versus bias in evaluation: Evidence from the NIH. *Amer. Econom. J. Appl. Econom.* 9:60–92.

March J (1991) Exploration and exploitation in organizational learning. *Organ. Sci.* 2:71–87.

Maslej N, Fattorini L, Perrault R, Gil Y, Parli V, Kariuki N, Oak S (2025) Artificial intelligence index report 2025. Accessed April 30, 2025, https://hai.stanford.edu/ai-index/2025-ai-index-report.

McDonald RM, Allen RT (2022) A spanner in the works: Category-spanning entrants and audience valuation of incumbents. *Strategy Sci.* 7:190–209.

McDonald RM, Eisenhardt KM (2020) Parallel play: Startups, nascent markets, and effective business-model design. *Admin. SSci.Quart.* 65(2):483–523.

Meincke L, Mollick E, Mollick L, Shapiro D (2025) Prompting science report 1: Prompt engineering is complicated and contingent. Preprint, submitted March 4, https://arxiv.org/abs/2503.04818.

Mintzberg H, Raisinghani D, Theoret A (1976) The structure of" unstructured" decision processes. *Admin. Sci. Quart.* 246–275.

Mollick E (2024) Reinventing the organization for GenAI and LLMs. *MIT Sloan Management Rev.*, 1–4.

Morris S, Oldroyd J, Allen RT, Chng DHM, Han J (2023) From local modification to global innovation: How research units in emerging economies innovate for the world. *J. Internat. Bus. Stud.* 54:418–440.

Murray A, Rhymer J, Sirmon DG (2021) Humans and technology: Forms of conjoined agency in organizations. *Acad. Management Rev.* 46(3):552–571.

Newborn M (2012) *Kasparov Versus Deep Blue: Computer Chess Comes of Age* (Springer Science & Business Media, New York).

Nickerson JA, Zenger TR (2004) A knowledge-based theory of the firm—The problem-solving perspective. *Organ. Sci.* 15:617–632.

Open LLM Leaderboard (2025) Retrieved September 15, 2025, https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

OpenAI (2024) Introducing OpenAI o1. Accessed April 15, 2025, https://openai.com/o1/.

OpenAI (2025) Introducing OpenAI o3 and o4-mini. Accessed April 30, 2025, https://openai.com/index/introducing-o3-and-o4-mini/.

OpenAI Pioneers Program (2025) Retrieved September 15, 2025, https://openai.com/index/openai-pioneers-program/.

OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. (2023) GPT-4 technical report. Preprint, submitted March 15, https://arxiv.org/abs/2303.08774.

Overview Leaderboard | LMArena (n.d.) Retrieved December 17, 2025, https://lmarena.ai/leaderboard.

Peterson A, Wu A (2021) Entrepreneurial learning and strategic foresight. *Strategic Management J.* 42:2357–2388.

Randazzo S, Lifshitz H, Kellogg KC, Dell'Acqua F, Mollick E, Candelon F, Lakhani KR (2025) Cyborgs, centaurs and self-automators: The three modes of human-GenAI knowledge work and their implications for skilling and the future of expertise. *Harvard Bus. Rev.*, https://www.hbs.edu/ris/Publication%20Files/26-036_e7d0e59a-904c-49f1-b610-56eb2bdfe6f9.pdf.

Rein D, Hou BL, Stickland AC, Petty J, Pang RY, Dirani J, Bowman SR (2024) GPQA: A graduate-level Google-proof Q&A benchmark. *Proc. 1st Conf. Language Modeling.*

Rivkin JW (2000) Imitation of complex strategies. *Management Sci.* 46(6):824–844.

Rivkin JW, Siggelkow N (2006) Organizing to strategize in the face of interactions: Preventing premature lock-in. *Long Range Planning* 39:591–614.

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, et al. (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587): 484–489.

UC Berkeley SkyLab (2025) LMArena. https://arena.ai/about.

Vinyals O, Babuschkin I, Czarnecki WM, Mathieu M, Dudzik A, Chung J, Choi DH, et al. (2019) Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575(7782): 350–354.

Wang Y, Ma X, Zhang G, Ni Y, Chandra A, Guo S, Ren W, et al. (2024) MMLU-PRO: A more robust and challenging multi-task language understanding benchmark. *Adv. Neural Inform. Processing Systems* 37:95266–95290.

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inform. Processing Systems* 35:24824–24837.

White C, Dooley S, Roberts M, Pal A, Feuer B, Jain S, Shwartz-Ziv R, et al. (2024) LiveBench: A challenging, contamination-free LLM benchmark. Preprint, submitted June 27, https://arxiv.org/abs/2406.19314.

Ryan T. Allen is an assistant professor of strategy at Brigham Young University's Marriott School of Business. He studies how organizations use data and artificial intelligence to make decisions that shape strategy and innovation. He received his PhD in business administration from Harvard Business School.

Rory M. McDonald is the John Tyler associate professor of business administration at University of Virginia's Darden School of Business. He studies how firms successfully navigate nascent industries and product categories. He received his PhD in management science and engineering from Stanford University's Technology Ventures Program.