Strategy Science

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

Can LLMs Aid Analogical Reasoning for Strategic Decisions? A Comparative Study

Prothit Sen, Maciej Workiewicz, Phanish Puranam

To cite this article:
Prothit Sen, Maciej Workiewicz, Phanish Puranam (2026) Can LLMs Aid Analogical Reasoning for Strategic Decisions? A Comparative Study. Strategy Science 11(1):118-136. https://doi.org/10.1287/stsc.2025.0426

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

STRATEGY SCIENCE
Vol. 11, No. 1, March 2026, pp. 118–136
ISSN 2333-2050 (print), ISSN 2333-2077 (online)

Can LLMs Aid Analogical Reasoning for Strategic Decisions? A Comparative Study

Prothit Sen,[a,*] Maciej Workiewicz,[b] Phanish Puranam[c]

[a] Department of Strategy, Indian School of Business, Hyderabad 500111, India; [b] Department of Management, ESSEC Business School, 95000 Cergy, France; [c] Department of Strategy, INSEAD, Singapore 138676, Singapore

*Corresponding author

Contact: prothit_sen@isb.edu, https://orcid.org/0000-0003-1409-1559 (PS); workiewicz@essec.edu, https://orcid.org/0000-0001-5745-719X (MW); phanish.puranam@insead.edu, https://orcid.org/0000-0002-0032-8538 (PP)

Received: May 1, 2025
Revised: July 21, 2025
Accepted: September 29, 2025
Published Online in Articles in Advance: February 10, 2026

https://doi.org/10.1287/stsc.2025.0426

Abstract. Analogical reasoning is central to strategy because it offers a basis for decision making in uncertain and data sparse contexts. Its effectiveness as a process depends not only on retrieving candidate analogies but on correctly matching them to the focal problem because a poorly chosen analogy can mislead decision makers and produce costly errors of commission. We investigate how humans and large language models (LLMs) perform at analogical reasoning through an exploratory study that extends classic analogical transfer designs by introducing multiple source analogs and target problems. Our results reveal a tradeoff: Humans in our sample frequently overlooked valid analogies (low recall) but rarely misapplied them (high precision); LLMs, in contrast, did not miss valid analogies (high recall) but often surfaced spurious, even if internally coherent matches (low precision). These findings suggest a complementary division of labor: LLMs might serve as expansive retrieval engines, generating a broad set of candidate analogies, whereas humans adjudicate their contextual fit through superior causal matching. This highlights a possible pathway for artificial intelligence (AI)–human collaboration in strategy making while underscoring the risks of over-reliance on AI-generated analogies until these models can improve their performance at matching analogies to problems.

History: Accepted for the Special Issue: Can AI Do Strategy?
Supplemental Material: The online appendix is available at https://doi.org/10.1287/stsc.2025.0426.

Keywords: AI-human collaboration • analogical reasoning • strategic decision-making • managerial cognition • matching complexity

1. Introduction

Analogical reasoning, where decision makers draw on solutions from familiar situations when dealing with novel challenges, is one of the central pillars of strategic thinking (Gavetti and Rivkin 2005, Miller and Lin 2015, Carroll and Sørensen 2024). Analogies serve as cognitive tools that help interpret complex or ambiguous situations by linking them to familiar precedents (Gentner et al. 2001). For example, when first being taught about atomic structure, it is common for students to be shown an analogy to the solar system, which exploits the fact that both involve smaller bodies revolving around a central attractor (Gentner and Smith 2013). Unlike other decision-making approaches that require relatively rich domain-specific information, analogical reasoning is particularly valuable in strategic contexts precisely because it enables inference and action when data are scarce. Analogies prove useful across a wide range of strategic choice situations, including decisions about market entry, acquisitions, business model innovation, and organizational turnaround (Gavetti et al. 2005). Beyond guiding choice, analogies also shape how managers mentally represent their strategic environment. By highlighting certain features and suppressing others, analogical reasoning influences how opportunities and threats are construed, thereby shaping the contours of strategic problem framing (Gary et al. 2012).

Access to a range of possible analogies is argued to be an advantage for strategists. When Charlie Munger famously remarked, "To become wise, you've got to have models in your head. And you've got to array your experience—both vicarious and direct—on this latticework of models" (as quoted in Page 2018, p. 1), he was noting the virtues of possessing a repository of analogies.[1] Yet his statement also highlights an important challenge: One must *select* the appropriate analogy from a broader set of possible candidates—a challenge known as the "matching problem" (Cummins 1992, Gentner et al. 1993). This is especially critical in strategy, where decisions are high stakes and often irreversible (Leiblein et al. 2018). Thus, a poorly matched analogy can lead to flawed inference, misplaced confidence, and costly strategic errors.

Without accurate matching, a large repository of possible analogies may, in fact, be a disadvantage.

Although stories of successful analogical transfer abound in the business world, it remains unclear whether these reflect the general power of analogies or simply the visible survivors, obscuring the many cases where analogical reasoning misfired. For instance, Kodak executives successfully modeled their film and camera business on Gillette's "razor-and-blade" pricing model, but also mistakenly treated digital images as analogous to traditional film, a mistake they did not recover from (Tripsas and Gavetti 2000). Similarly, Wayne Huizenga successfully leveraged his experience consolidating funeral homes to turn local video rental businesses into Blockbuster. Yet Blockbuster ultimately faltered by treating Netflix's business model as analogous to, and therefore redundant with, its own DVD distribution system. Overall, how decision makers identify good matches and filter out poor analogies remains an important and yet poorly understood phenomenon (Blanchette and Dunbar 2001, Gentner and Smith 2013).

Recent advances in artificial intelligence, and in large language models (LLM) in particular, offer a fresh perspective on this question. LLM technologies are directly relevant to the challenge of effective analogical reasoning. At their core, the transformer architectures underlying these models (Vaswani et al. 2017) enable them to detect similarity and relevance between ideas expressed as text—that is, between a given prompt and patterns embedded in their training data. However, although similarity detection is necessary, it is not sufficient for effective analogical reasoning (Olguín et al. 2022). Even if LLMs can outperform humans in retrieving potential analogies based on similarity, their ability to generate high-quality matches remains unproven. On the dimension of structural abstraction, particularly in the domain of verbally formulated reasoning, which is common in business contexts, there is scant evidence comparing the performance of LLMs and humans (Yuan et al. 2023; for nonverbal reasoning, see Lewis and Mitchell (2024) and Camposampiero et al. (2025)). Put differently, can LLMs both retrieve and correctly match analogies to the problems at hand, and how does their performance compare with that of humans? Moreover, can human decision makers benefit from analogical reasoning executed by LLMs? What complementarities in analogical reasoning exist between humans and AI?

This paper directly investigates these questions by comparing the verbal analogical reasoning abilities of modern LLMs and humans in the context of business problems. In a novel exploratory study designed to meaningfully extend classic work on analogical transfer (Gick and Holyoak 1980, 1983), we introduced a matching problem by pairing two source analogs with two target problems that reflect typical business contexts, but which have well-defined causal schemas. Human participants ($n = 199$) and a set of eight LLMs (total of $n = 512$ trials) were tasked with solving each target problem under both "no hint" and "hint" conditions, where a hint made the existence of a possible analogical frame explicit (without pointing out the correct one). *Correctness* of the solution was measured by whether the individual respondent solved the problem using the correct underlying causal schema, whereas *matching accuracy* was measured at the sample level as the fraction where a correct analogy was used and incorrect ones were discarded in the set of all responses to the problem-solving task.

We find that the latest generation of reasoning models rival humans in correctly solving the problems we posed, although performance varies considerably across models. With the aid of a hint, reasoning models solve the problems correctly in 60%–80% of cases compared with 10%–30% for human subjects. However, when it comes to matching accuracy, our results reveal a precision–recall tradeoff between humans and LLMs in analogical transfer. Humans demonstrate high precision—making few false matches (0.70–0.75 with a hint)—but low recall, often failing to detect true matches when they occur (0.20–0.40). LLMs, by contrast, achieve consistently high recall (often near-perfect, ~1.0 with a hint), surfacing a broader set of candidate analogies. Without a hint, LLMs generally underperform humans on precision, although reasoning LLMs approach human levels of precision when cued. Further, the errors of commission these reasoning LLMs commit reflect internally coherent but misaligned solutions, posing the risk of producing convincing but inappropriate analogical matches.

This insight into the misapplication of analogy by LLMs is critical. In strategic problem solving, applying a false analogy may be at least as, if not more, damaging as overlooking a good one. Misapplied analogies, such as Kodak's treatment of digital and film-based imaging as equivalent, or Blockbuster's view of Netflix as analogous to DVD rental, can lead to costly business failures. Additionally, leaders who espouse faulty analogies risk losing legitimacy with subordinates, who may lose confidence in their vision. The explainability and persuasive force of strategic vision is arguably as important, for its coordinative function, as its validity (Koçak and Puranam 2022), and false analogies undermine both. By contrast, leaders who miss analogies may still be able to benefit from others in their team being able to discover them, and if none does, the error often remains invisible. As in many business settings, errors of commission are often more visible and expensive than errors of omission (March 1991, Csaszar 2013), and analogical reasoning may be no exception.

In terms of managerial relevance, our findings suggest a complementary division of labor: In their

current state of development, LLMs function best as expansive analogy-retrieval engines, whereas human managers provide the precision necessary to adjudicate which causal schemas genuinely fit the focal problem context. This synergy highlights a pragmatic pathway for AI–human collaboration in strategic problem solving. Future generations of LLMs will doubtless surpass current capabilities, and their relative superiority in generating a broad set of candidate analogies is likely to persist. It remains an open question whether their capacities at matching will eventually exceed those of humans (Olguín et al. 2022, Yuan et al. 2023); but until they do, human involvement in analogical reasoning will be necessary for successful analogical transfer to aid strategic decision making in organizations.

2. Analogy and Strategy: Theoretical Background

Analogical reasoning is of unique interest in strategic decision making because it provides a mechanism for inference under conditions of *low data and high uncertainty*—facets of incomplete information that pose particular challenges for boundedly rational managers (Simon 1997). Unlike other cognitive processes, such as case comparison or search and evaluation that presuppose a relatively rich set of domain-specific data, analogical reasoning allows decision makers to generate plausible hypotheses and courses of action when such data are unavailable. In this sense, analogies function as a form of abductive reasoning: they offer provisional explanatory or predictive schemas that can guide both hypothesis testing and action in novel environments (Gavetti et al. 2005). More broadly, they are a particular form of generalization, where knowledge about what has been experienced is assumed to apply to what has not yet been experienced (Choi and Levinthal 2023, Levinthal and Schliesmann 2025). By linking the unfamiliar to the familiar (Carroll and Sørensen 2024), analogies help managers navigate complexity and act as bridges between theory and practice, cognition and action. This property makes analogical reasoning particularly salient in strategy, where managers must often act under ambiguity, and where the value of an analogy lies not only in its correctness but in its ability to frame the problem and generate testable alternatives.

Strategy scholars have theoretically addressed analogical reasoning in a variety of organizational contexts (Miller and Lin 2015), and implicit in this body of work is the underlying process of abductive reasoning under incomplete information. It has been argued that managers frequently draw analogies from past experiences to frame new challenges (Tripsas and Gavetti 2000, Gavetti et al. 2005). Analogies also shape how managers perceive opportunities and threats (Gary et al. 2012). In entrepreneurial settings, analogies serve to frame new ventures, drawing on familiar categories that guide investor expectations and strategic alignment (Santos and Eisenhardt 2009, Navis and Glynn 2011). Similarly, analogical reasoning facilitates capability reconfiguration in dynamic markets by transferring lessons from prior contexts, sometimes from domains as disparate as wars and sports, to novel contexts (Zollo and Winter 2002, Helfat and Peteraf 2003, Cornelissen et al. 2011). As such, analogical reasoning and associated pattern recognition has been argued to allow managers to cope with rapid change and adaptation (Gavetti et al. 2005).

Within strategic practice, analogies serve two main purposes: generative and rhetorical. Generatively, analogies function as cognitive tools for problem solving, enabling managers to transfer insights from past experiences or other industries into new, often ambiguous contexts (Gavetti and Rivkin 2007, Gary et al. 2012). Gavetti and Rivkin (2007) identified analogical thinking as a powerful element of strategic search that complements local experimentation, deductive reasoning, and imitation. The strength of analogical reasoning lies in its ability to simplify complexity, frame ill-structured problems, and spark creativity. It can generate breakthrough insights by highlighting structural similarities across seemingly unrelated domains, a process at the heart of visionary strategies and creative leaps (Brandenburger 2017). Analogical thinking is also the key mechanism that allows organizations to bring new and diverse knowledge from outside and integrate it into their own knowledge stock and processes (Tschang and Ertug 2016). For instance, analogies can help generate new business models, as in the cases of JustPark, ParkLet, or SpotPog that applied the Airbnb model to rent unused parking spaces in cities (Schilling 2018).

Rhetorically, analogies act as persuasive devices, allowing strategists to communicate ideas vividly and rally support by drawing parallels to recognizable domains such as sports, war, or ecosystems (Carroll and Sørensen 2024). For example, using analogies from sports or comparing organizations to a well or poorly oiled machines is often employed in internal communications (Duhaime and Schwenk 1985). This dual role makes analogy an indispensable part of the strategist's toolkit, especially when deductive reasoning is too demanding or when data are incomplete (Carroll and Sørensen 2024).

However, the weakness of analogical reasoning is equally significant: Analogies often emphasize surface resemblances rather than deeper causal structures, potentially biasing choices and narrowing the range of considered options (Gavetti et al. 2005, Gary et al. 2012). Poorly chosen analogies may lead to overconfidence,

misapplied lessons, or flawed decisions such as ill-conceived acquisitions or divestments (Duhaime and Schwenk 1985). As research shows, their effectiveness depends on how managers build and evaluate them: Multiple analogies, structural alignment, and exposure to variation can improve transfer and reduce bias (Gary et al. 2012). In this sense, analogical reasoning is both a source of strategic creativity and a potential trap, making it an essential yet double-edged cognitive process that shapes the origins, formulation, and communication of strategy (Brandenburger 2017, Schilling 2018).

2.1. Retrieval, Mapping and Matching: Foundations of Analogical Reasoning

Research in cognitive psychology recognizes analogical reasoning to be a fundamental cognitive process that enables the transfer of knowledge across domains (Gentner 1983, Holyoak and Thagard 1989). Researchers have described two subprocesses as foundational to analogical reasoning: retrieval, where possible source analogies come spontaneously to mind, and mapping, where the candidate analogies are compared with the target problem at hand. Underlying both subprocesses are the concepts of mental representations and similarity between them (Holyoak and Thagard 1995).

One can represent two problem domains using graphs (networks) that capture the key concepts and the causal interconnections between them. Let the true causal structure in a domain be the graph G. The decision maker learns to represent this as g. In general, g may differ from G in nodes, links, or both; that is, representations can be imperfect. Let the true causal structure in the target domain (where we want to apply analogical insight from other source domains) be G_T. The decision maker learns to represent this as g_T. The problem, therefore, for analogical reasoning is to find a source domain i such that g_i is a better approximation than g_T (as well as a randomly generated g_R) of G_T.

Two graphs may resemble each other in terms of the nodes used (superficial or surface similarity) or the pattern of relationships between the nodes (structural or deep similarity). Successful analogical reasoning depends on finding structural similarities between situations, so that knowledge of a solution in one situation (source) can act as a useful hypothesis about the second (target) (Gentner and Smith 2013, Goldwater and Gentner 2015). Structural mapping involves checking for structural similarity across source and target domains (Gentner 1983) via the detection of correspondence between "causal relations in the two situations" (Holyoak and Koh 1987, p. 334). In graph terms, the semantic similarity constraint in Holyoak and Thagard (1995) corresponds to node overlap between g_i and g_T, the structure constraint corresponds to overlap in the pattern (topography) of links, and the purpose constraint reflects whether the selected g_i is a better approximation than g_T of G_T.

For human decision makers, analogical retrieval is often based on superficial similarity (Gick and Holyoak 1980). Each feature in the source domain "activate[s] memory representations of other situations that share that feature" (Holyoak and Koh 1987, p. 333). Training that emphasizes the detection of structural alignment over superficial resemblance improves the quality of analogical reasoning (Gick and Holyoak 1980). Research also demonstrates that comparative exposure to analogs sharpens structural similarity detection (Vendetti et al. 2015, Richland and Begolli 2016, Lovett and Forbus 2017). Prior knowledge and abstraction ability also improve schema recognition and analog selection (Hofstadter 2001, Goldwater and Gentner 2015).

When multiple candidate analogies g_i are retrieved, beyond mapping each to g_T, there is the additional challenge of comparing across these pairs to detect the candidate with the best structural alignment to the target situation. Decision makers can fail this challenge if they satisfice and stick with the first satisfactory analogy or fail to compare the alternatives based on structural similarity (Gentner and Smith 2013). In other words, mapping governs how well a given candidate analogy fits; matching governs which candidate is selected when several are retrieved with varying degrees of correspondence arrived at through mapping. Matching thus involves the ability to conduct mapping on each retrieved candidate and compare the strength of mapping across candidates on the dimension of structural similarity.

To illustrate these concepts—retrieval, mapping, and matching—consider a midsized drone manufacturer that is currently evaluating a strategic shift from outright product sales to a Drone-as-a-Service model. Under this proposed model, clients pay per-flight-hour, whereas the drone manufacturer retains ownership and handles maintenance. The most readily accessible analogy—the one that is *retrieved*—may be the widely recognized "razor-and-blade" pricing model. Both involve durable goods complemented by a stream of recurring revenues, rendering this example cognitively salient due to surface similarity. However, a more careful examination of the relational structure reveals critical misalignments: in the razor-and-blade model, vendors generate margin through the sale of consumables, whereas the drone maker would most likely treat maintenance and spare parts as cost centers while monetizing equipment availability. A less likely to be retrieved but *structurally* superior analogue is the "power-by-the-hour" model pioneered by Rolls-Royce in the context of jet engines, where the manufacturer retains ownership, earns revenue in proportion to usage, and bears operational risks traditionally held by the customer. It represents a better match than the razor-and-blade analogy,

in terms of similarity of the pattern of causal relationships. Both analogies (razor-and-blade and power-by-the hour) must be *retrieved*, each must be *mapped* to the drone manufacturer's situation, and one must be selected based on superior *match*.

Put simply, analogical reasoning can fail if good analogies are not retrieved, and among the retrieved candidate analogies, matching may fail if decision makers satisfice and stick with the first satisfactory analogy rather than iterate through alternatives to assess structural similarity (Gentner and Smith 2013). Although cognitive scientists are fully aware of the challenges of matching (Forbus et al. 1995), studies that explicitly set up a matching problem are rare (Gentner and Smith 2013). Laboratory studies (Gick and Holyoak 1983, Cummins 1992, Markman and Gentner 1993) typically hand participants a lone source, so they bypass the need to choose among several alternatives. Naturalistic observations (Blanchette and Dunbar 2001) record which analogies were selected but cannot measure whether *better* analogies were available or how accurate the choices were.

The matching problem is also well recognized by theorists of analogical reasoning in economic contexts. The analogy-based expectation equilibrium of Jehiel (2005) formalizes how agents simplify complexity by grouping situations into "analogy classes" and forming expectations at that level of abstraction. The challenge is to choose the most relevant class. Carroll and Sørensen (2024) emphasize that the central challenge of analogical reasoning in strategy is not whether analogies are used (they are unavoidable given low data and high uncertainty in decision situations) but whether the *right* ones are selected. The matching problem, they propose, arises because surface resemblances are easy to detect, whereas structural causal alignments are harder to discern yet decisive for strategic relevance.

2.2. Analogical Reasoning by Machines

Contemporary approaches to analogical reasoning in machines have evolved significantly, shifting from early rule-based systems toward data-driven methods capable of identifying analogies from large-scale datasets (Pan et al. 2010, Sun and Saenko 2016). Techniques such as transfer component analysis and correlation alignment (CORAL) aim to minimize statistical divergence between domains by aligning feature distributions, whereas more recent innovations, including probabilistic analogical mapping, enable forms of zero-shot reasoning by leveraging structural inferences (Webb et al. 2023b). In parallel, developments in causal modeling frameworks have foregrounded the importance of relational over surface-level similarity, emphasizing structural isomorphism in analogy construction (Correa et al. 2022, Pearl and Bareinboim 2022).

Most current applications of LLMs to analogical reasoning have focused on highly stylized tasks. These include pattern completion problems such as nonvisual Raven's matrices, digit sequence analogies, molecular property transfer, and multiple-choice science questions—contexts in which a single source-target pair is prespecified, and the model is tasked only with inferring a missing element or selecting among fixed alternatives (Webb et al. 2023a, Yuan et al. 2023, Lewis and Mitchell 2024). Although such benchmarks provide useful diagnostics of a model's ability to induce structural patterns (e.g., via structure mapping or abduction), they do not capture the open-ended, linguistically framed analogical tasks that characterize real-world managerial reasoning. Questions such as whether a proposed Drone-as-a-Service model more closely resembles a razor-and-blade pricing scheme or a jet engine leasing arrangement exemplify the type of analogical judgment that existing evaluations bypass.

Equally significant is the neglect of the matching problem in current studies. In many settings, candidate analogues are either preselected or generated through retrieval mechanisms whose outputs are not subjected to any formal evaluation process. As a result, we lack systematic evidence on whether and how well LLMs can identify the most structurally appropriate source when multiple plausible analogies are available—a decision step that is critical in high-stakes reasoning.

In this study, we address these gaps by designing an experimental task that (i) employs naturalistic business vignettes written entirely in free-form language, (ii) embeds an explicit two-analogue matching decision, and (iii) evaluates performance across the full analogical reasoning pipeline—*retrieval*, *mapping*, and *matching*—for both humans and LLMs.

3. Study Design

We designed an exploratory comparative study to investigate how humans and artificial intelligence (AI) systems navigate the matching problem in analogical reasoning. Unlike confirmatory studies that test specific hypotheses, our exploratory approach allowed us to discover patterns in how different cognitive systems (human and LLM-based AI) handle competing analogies. Although we preregistered the replication of the "hint effect" of Gick and Holyoak (1983) for validation purposes, the primary focus of our study is to explore and uncover the differences in matching behavior and efficacy of analogical transfer between humans and LLMs.

3.1. Setup

The influential studies by Gick and Holyoak (1980, 1983) demonstrated that when participants were

given a single source story (e.g., a general attacking a fortress) and asked to solve a target problem (e.g., a doctor treating a tumor), many could apply the source, especially when given a hint. But this one-to-one setup does not involve matching, because it does not test whether people can discriminate between multiple, competing potential analogies—some superficially similar, others structurally appropriate. In other words, the Gick and Holyoak studies illuminated how analogical *use* can be elicited (i.e., they are tests of retrieval and mapping), but not whether analogical *matching* can be competently performed under realistic cognitive conditions. In our studies, we incorporated a matching problem, where participants were shown two source stories and faced two target problems. Our experiments thus departed from the classical "Radiation problem" studies on two important aspects as follows:

i. **Introducing matching complexity**: In addition to the Radiation problem story (Story 1), we introduce an additional story in the source domain (Story 2). Mirroring this in the target domain, we expose the subjects to two problems (Problem 1 and Problem 2) instead of one. To solve Problem 1, Story 1 should serve as the appropriate analogy, whereas Story 2 would act as a placebo. Similarly, to solve Problem 2, Story 2 should serve as the appropriate analogy, whereas Story 1 would act as a placebo.

ii. **Humans versus AI (LLMs)**: We run the above setup for human subjects and independent trials on several state-of-the-art LLMs (GPT-4, DeepSeek V3, Gemma 3 27B, Llama 3.3 70B, Mistral 7B Instruct, Gemini 2.5 Flash Lite, Grok 3 mini, and o3-mini). The human subjects comprise Master-level students from a leading business school.

3.2. Source Stories and Target Problems

We construct an experimental setup with two source stories (S1 and S2) and two target problems (P1 and P2). Following Gick and Holyoak, we predetermine a "correct analogical mapping" (between S1 → P1 and S2 → P2) based on overlaps in the underlying causal schemas between the source stories and target problems. However, participants are not informed of these mappings, which creates the possibility of incorrect pairings (e.g., S1 → P2 or S2 → P1).

The target problems are described through stylized vignettes that are not intended as full industry cases but to isolate the core challenge of matching. Stylization lets us evaluate what matters for strategy analogies: identifying premises, testing *negative* premises, and, crucially, checking vertical (causal) rather than purely horizontal (surface) similarity (Carroll and Sørensen 2024). In this sense, our vignettes are model-cases (Gilboa et al. 2014) that probe whether decision makers place the focal problem in the right analogy class (Jehiel 2005) and thereby build a valid strategic frame (Rindova and Martins 2023).

We retain S1 (the Radiation problem) from the original stories of Gick and Holyoak (1983) to benchmark our results against prior findings. The corresponding target problem, P1, is a novel but structurally analogous scenario in the domain of operations and supply chain management, where the common causal schema for S1 and P1 is "split and converge." We design S2 and P2 as a new pair, where mapping relies on recognizing *survivorship bias* as the abstract causal schema underlying both. This bias is broadly applicable across managerial contexts, especially when making inferences from incomplete data. The detailed texts of the two stories (S1, S2), and the two problems (P1, P2) are provided in the Online Appendix (Section 1).

Both problems capture, albeit in a stylized form, widely recurring problems that strategists face. The challenge of managing economic development under sustainability and conservation constraints is widespread (Battilana et al. 2022), as is the problem of best practice implementation without adequate accounting for survivorship bias (Denrell 2003). The two stories (about radiation and dolphins) contain causal schemas that, if recognized, matched, and applied to the correct problem, can suggest a solution. These causal schemas, and how they apply to the problems, are illustrated in Figure 1, (a) and (b).

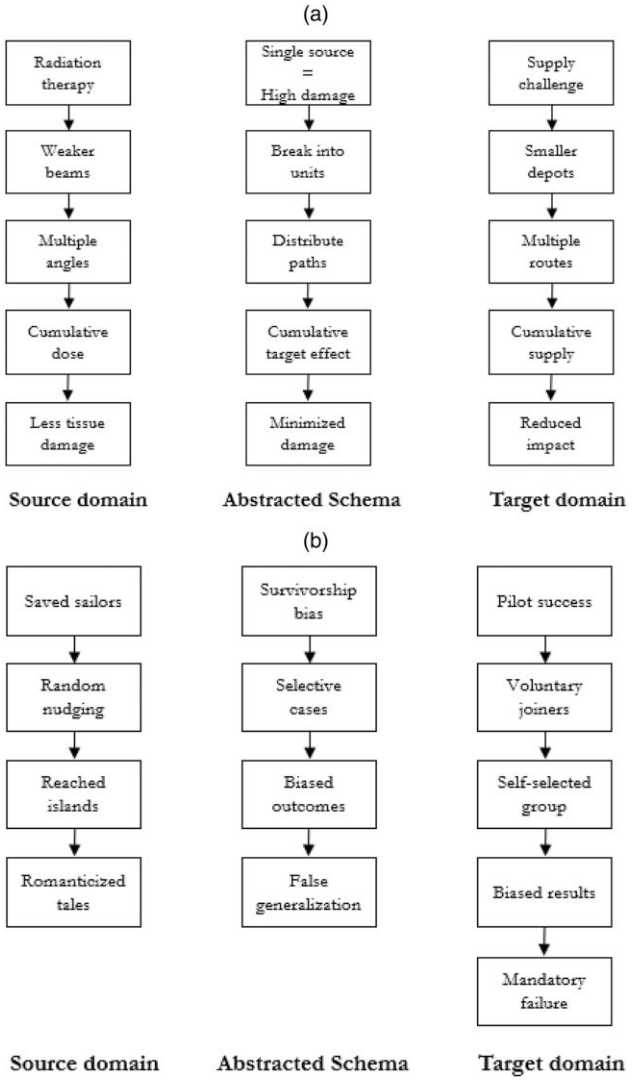
3.3. Protocol for Participants

For our human participants, we follow the same protocol as Gick and Holyoak (1980, 1983) and begin by verifying that participants apply the analogy more often under hint than under no hint conditions. This study was preregistered (https://aspredicted.org/rsb2-xgmx.pdf). This replication strengthens the validity of our protocol, which includes new stories and problems beyond those in Gick and Holyoak (1983). We lead our participants to solve the problems in the target domain in the stepwise process.

Exposure to stories: We begin by providing the participant with each story in turn. We instruct the participant to read the stories carefully, summarize them, and rate the story's *ease of understanding* and *plausibility* using a five-point Likert scale (Gick and Holyoak 1983). Ease and plausibility were evaluated on a five-point Likert scale, from one (least easy/plausible) to five (very easy/plausible).[2] We allow three minutes per story and randomize the order in which the stories are presented.

No hint condition: We present the participant with a problem-solving task, with their goal being to generate as many solutions as possible. The subjects are randomly allocated to either the Factory problem or the HR problem. We allow the participants six minutes to

Figure 1. Source Story, Causal Schema, and Target Problem



Notes. (a) Causal schema for Radiation story to city Factory problem. (b) Causal schema for Dolphin story to HR problem.

solve the problem and ask them to write down their solutions as text (we provide them the option of providing multiple solutions in multiline text box).

Hint condition: In the instructions for the problem-solving task, we explicitly offer the following hint: "You can use one of the stories from the first text summarization exercise to solve the puzzle."

LLMs: We follow the same sequence for each independent run of each LLM model. Initially, we conducted our analysis with a single LLM, OpenAI's GPT4. Because this was exploratory, it was not preregistered. Subsequently, we replicate our analysis with data from seven other state-of-the-art LLMs, and this study was preregistered (https://aspredicted.org/m992-r873.pdf). For LLMs, we do not have to monitor time to read and summarize the stories or generate a solution to the problems as we do for the human participants.

Following each problem-solving task, we explicitly ask the models whether they had used either of the preceding stories in their solutions and which story they considered more useful, allowing us to assess both spontaneous analogical transfer and declared "awareness" of analogical reasoning processes.

3.4. Sample

We conducted our study using human and AI subjects (LLMs). Human subjects ($n = 210$) were undergraduate students taking the core strategy class in a leading business school based in France. After applying the preregistered attention check (minimum time to complete the survey $t_{\min} = 3$ minutes and including only fully completed surveys) and removing two invalid responses (two subjects pasted their answers from an LLM, even though they were explicitly asked not to do so), we excluded 11 participants. The final sample consisted of $n = 199$ participants: with 101 (50.8%) male, 93 (46.7%) female, and 5 (2.5%) subjects who did not indicate their gender. The study was approved by the School's Ethical Review Board, and all participants provided informed consent.

We evaluated analogical reasoning across eight LLMs using a within-subjects experimental design but otherwise replicated the procedure used with human participants. The models selected for the study were GPT-4 (OpenAI 2023), DeepSeek V3 (DeepSeek-AI 2024), Gemma 3 27B (Google DeepMind Team 2025), Llama 3.3 70B (Dubey et al. 2024), Mistral 7B Instruct (Jiang et al. 2023), Gemini 2.5 Flash Lite (Google DeepMind Team 2025), Grok 3 mini, and o3-mini (OpenAI 2025). All of the above are transformer-based LLMs and were queried through APIs using OpenRouter server (OpenRouter 2025). To replicate the default settings of these models, we use a temperature of 0.7.

These models represent distinct approaches to creating large language models. Three of these models (Gemini 2.5, o3-mini, Grok 3) are reasoning models; that is, they employ deliberative chain-of-thought processes during inference to solve complex problems through step-by-step logical analysis, typically spending additional computational time to "think" before generating responses, which distinguishes them from standard language models, like DeepSeek V3, GPT-4, Gemma 3, Llama 3.3, and Mistral 7B, which generate tokens sequentially without explicit intermediate reasoning steps (OpenAI 2024). Models like o3-mini and DeepSeek V3 are more powerful, with the latter having 671B parameters (OpenAI does not disclose the number of parameters for their o3-mini model), whereas Mistral 7B has only 7 billion parameters. This approach provides a representative sample of modern LLMs.

We used these LLMs to explore the generalizability of matching patterns across different AI architectures. Each model "solved" Problem 1 (Factory) and Problem 2 (HR) under both hint and no hint conditions. For each model, we conducted 16 separate trials per experimental condition (defined by the combination of hint/no hint and problem 1/problem 2), with each trial freshly reinitiated from a new state to prevent context carryover across trials. This produced a total of 512 responses. Given the relative stability observed in AI-generated outputs to identical prompts across models, a cell size of 16 observations per condition was sufficient for the purposes of this analysis (Yasunaga et al. 2023).

3.5. Measuring Correctness and Matching Accuracy

For measuring *correctness* of solution at the level of the individual respondent, we created an objective measure of semantic or causal overlap, automated via an "impartial coder"—GPT-5 (a state-of-the-art LLM that was *not* among the experimental subject LLMs). GPT-5 was tasked with computing the degree of structural similarity between the causal schemas underlying the two source stories (Figure 1, (a) and (b)) and the responses to the two problems through a semantic-embedded similarity measure.

The process of computing this measure involves the encoding of text into numerical vectors (embeddings) and then calculating the distance between these vectors to assess how semantically related two texts are, capturing similarity in meaning rather than surface form. This formulation incorporates both the extent of overlap (i.e., the fraction of causal keywords matched) and the ordering consistency between the causal schema and the proposed solution. For instance, in the HR problem, if "False Generalization" appears after "Biased Outcomes," its contribution to the similarity score is weighted more heavily, because this ordering reflects its downstream role in the causal chain (with respect to the Dolphin story's schema). The detailed mathematical formalization and operationalization of the semantic-embedded similarity score by GPT-5 is provided in the Online Appendix (Section 2). In the main analysis, we report results using the medium-threshold measure of structural overlap, although the qualitative insights remain unchanged when applying either the stringent or the lenient thresholds.[3]

Measuring *matching accuracy* at the aggregate sample level entailed calculating the proportion of cases in which respondents applied the correct analogy while rejecting incorrect ones, relative to the total number of respondents in a given condition. This was operationalized through a detailed confusion matrix analysis, which distinguished between claimed analogy use by the respondent and whether that analogy is actually the right analogy to solve the problem; the full procedure is described in the next section.

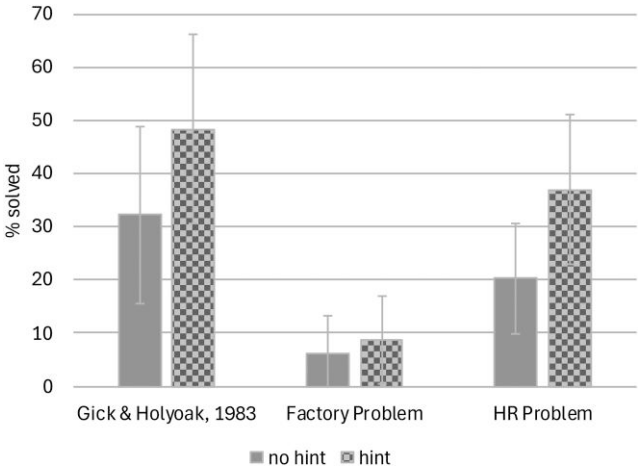
4. Results

4.1. Solution Correctness

We first measure Correctness at the individual subject level, capturing whether the correct underlying schema was applied to solve the problem. Importantly, this measure does not, by itself, establish that an analogical transfer occurred: A subject may arrive at the correct causal schema either by drawing on the appropriate source story or independently (without necessarily invoking analogy from the source story).

4.1.1. Human Subjects. Human subjects found the correct solution—that is, used the appropriate underlying causal schema—in solving the HR problem (survivorship bias schema) more readily than the Factory problem ("split and converge" schema) in both hint and no-hint conditions. Second, as expected, providing a hint to use an analogy increased its use for both problems. Although 8.7% (4 of 46) of respondents solved the Factory problem after being given a hint (versus 6.3% without a hint), nearly 37% (17 of 46) of respondents solved the HR problem after a hint (versus 20.3% without a hint). Third, correctness on both the Factory and HR problems was lower than on the original problems in the classical Gick and Holyoak (1983, experiment 2) study, where 29% (8 of 28) of respondents solved the problem without a hint and 50% (14 of 28) with a hint. These results (Figure 2) reveal two important insights. First, the solvability of problems (with or without a hint) reflects elements of problem- or context-specificity. In this case, survivorship bias was easier to spot than the "split and

Figure 2. Correct Solutions from Human Subjects



Note. Whiskers denote the 95% confidence interval; % solved denotes proportion of all responses in which the correct causal schema was applied to solve the problem.

converge" schema relevant to the Factory problem. Second, the lower overall success rates at correctly solving the problems compared with the original studies suggest the importance of the matching problem. The mere fact that there was an element of matching from two source stories to two target problems, compared with a one-to-one mapping in the original studies, may partly account for the differences between this and the original studies.

4.1.2. LLM. First, similar to the case of human subjects, the probability that LLMs provide a correct solution to the problem increases when provided with a hint (60%–80% with hint versus 10%–50% without hint). Second, reasoning-oriented LLMs outperform both nonreasoning LLMs and humans in the task of providing the correct solution (e.g., models like Grok-3 mini achieve 100% correctness with hint). Third, heterogeneity is evident within LLM performance itself: Some problem types prove more tractable than others. For instance, in the present study, LLMs demonstrate a clear facility in recognizing and applying the survivorship bias mechanism, even in the absence of hints. The case of GPT-4 is particularly revealing. Although formally categorized as a nonreasoning, or at best pseudo-reasoning model, its performance aligns closely with that of reasoning models. Conversely, when the underlying schema is relatively complex—as is arguably the case in the Factory problem—most nonreasoning LLMs (with the exception of DeepSeek) exhibited poorer performance; even with hints, their performance gains remain limited (correctness of 10%–50% with hint, compared with 20%–80% for the HR problem).

In sum (Figure 3, (a) and (b)), reasoning-based LLMs, particularly when aided by hints, achieve successful solution rates of 60%–100%, thereby exceeding the performance of human subjects, whose solution rates range between 10% and 35% in our study. Strikingly, even under matching complexity (i.e., the experimental setting of our study), reasoning LLMs approximate the performance levels reported in the original Gick and Holyoak studies (approximately 50% solvability) even though those studies did not incorporate the additional challenge of matching. This represents an important frontier in decision-making research, underscoring the convergence of reasoning LLMs with human-level abstraction capabilities in strategic problem solving.

4.2. Matching Accuracy: Precision vs. Recall

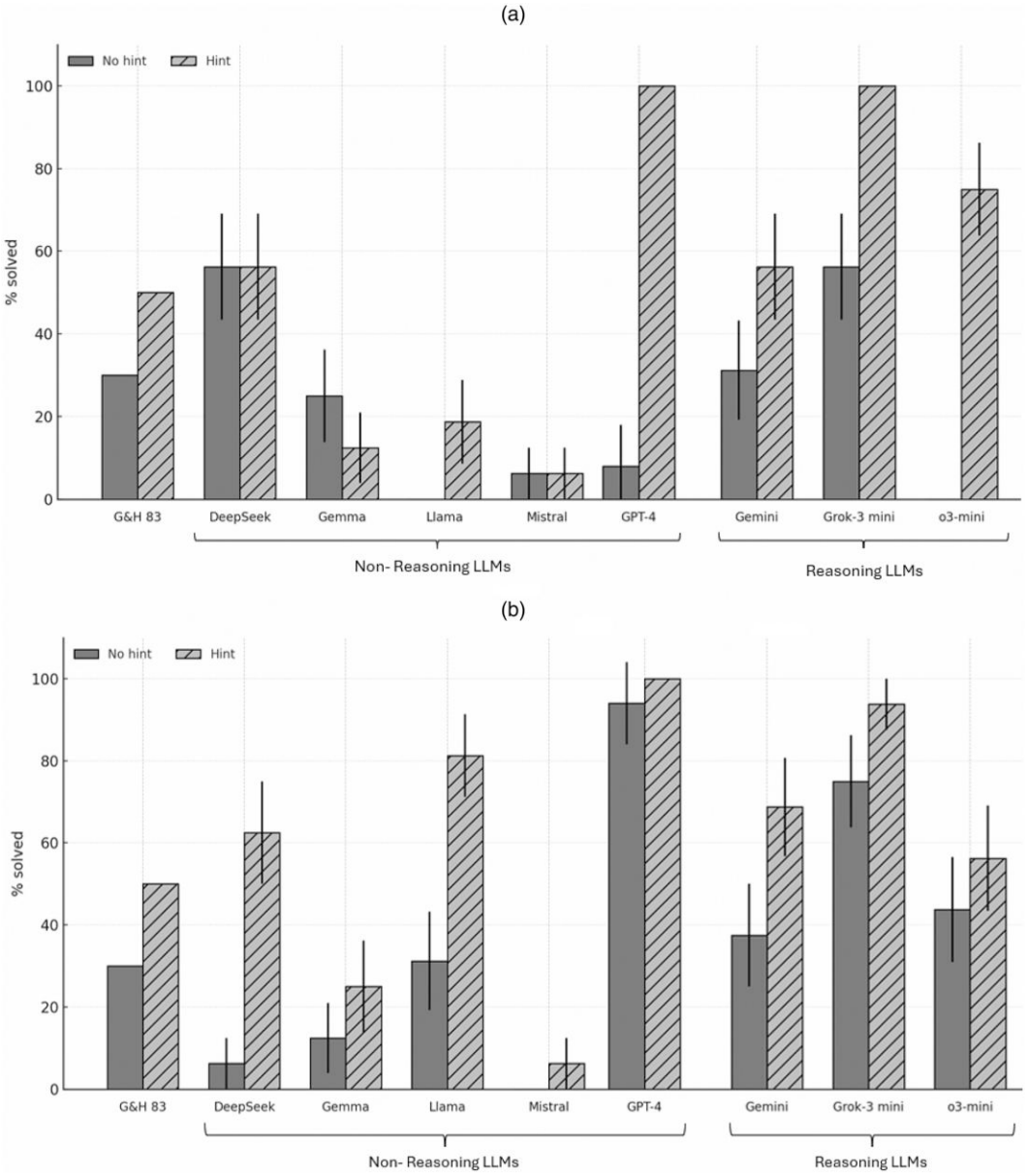
Although it is possible to solve a problem correctly without having applied an analogy, applying an analogy to solve a problem does not, by itself, ensure success, particularly when multiple candidate source analogies are available: This is the matching problem. The critical question is whether the solver can both retrieve a potentially relevant analogy and accurately match its underlying causal schema to the target problem (i.e., also discard inappropriate ones).

To disentangle these two mechanisms, we analyze the matching accuracy in both human and AI responses using a confusion matrix framework. We consider performance in two conditions: when subjects were not provided with a hint and when they were explicitly cued to use an analogy. We construct confusion matrices defined to record whether a given story represented the correct source analogy for the problem and the number of the responses that employed it (claimed to have used it).[4] For example, Table 1 displays the confusion matrices for human subjects in the "no hint" condition. The goal of this analysis is thus to isolate whether matching accuracy of analogical transfer arises primarily from recall (the retrieval of a potential analogy) or from matching (the alignment of the correct causal schema). From these confusion matrices, we get the distribution of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) (Table 2). In our specific experimental context, the response to a given problem was recorded as TP when the respondent claimed to use a specific reference story and the answer to the problem was indeed the causal schema from that story. A response was recorded as FP if the respondent claimed to use a specific story but the causal schema underlying that story was not the correct one to solve the problem. A response was recorded as FN if the respondent did not claim to use the specific reference story but that story indeed contained the correct causal schema to solve the problem. Finally, a response was recorded as TN if the respondent did not claim to use the specific reference story, and indeed that story did not have the correct causal schema to solve the problem.

We construct similar confusion matrices for both possible analogies, under both the no hint and with hint conditions and for both human subjects and all LLM models. From these matrices, we derive three standard metrics that capture distinct dimensions of an analogical match. As per standard practice, *Precision* is defined as the ratio of true positives to the sum of true positives and false positives, thereby indicating the ability of an agent class to identify correct analogical matches relative to all attempted matches. *Recall* is defined as the ratio of true positives to the sum of true positives and false negatives, capturing the extent to which correct analogical matches are successfully retrieved. Finally, *Matching Accuracy* is defined as the proportion of correctly classified cases—true positives and true negatives—relative to the total number of responses in the matrix.

Importantly, distinguishing whether overall matching accuracy is primarily driven by precision or recall provides insight into the underlying mechanism of problem solving through analogical transfer. A high

Figure 3. Correct Solutions from LLMs



Notes. (a) Correct solutions from LLMs for city Factory problem. (b) Correct solutions from LLMs for HR problem. Whiskers denote the 95% confidence interval (CI); CI for the LLMs should be treated with caution as the responses are not i.i.d; % solved denotes proportion of all responses in which the correct causal schema was applied to solve the problem; and missing bars suggest that the % solved for those particular conditions is zero.

Table 1. Confusion Matrices

Panel A: Confusion matrix for Radiation story (human agents)		
	Claimed to use Radiation story	Did not claim to use Radiation story
Truth: Radiation story is the right analogy	3	45
Truth: Radiation story is not the right analogy	4	55
Panel B: Confusion matrix for Dolphin story (human agents)		
	Claimed to use Dolphin story	Did not claim to use Dolphin story
Truth: Dolphin story is the right analogy	13	46
Truth: Dolphin story is not the right analogy	0	48

Table 2. Confusion Matrix Summary for No-Hint Condition of Human Subjects

Case	TP	FN	FP	TN
Human: Radiation	3	45	4	55
Human: Dolphin	13	46	0	48

Precision score suggests that the class of agents (human or LLM) excels at matching the correct causal schema to the target problem, whereas a high Recall score reflects better retrieval of potential analogies—a process emphasized in the original studies of Gick and Holyoak (1983). Figure 4, (a)–(c), presents the resulting Precision, Recall, and Matching Accuracy scores across all experimental conditions.

4.2.1. Precision. Significant contrasts between humans and LLMs emerge in the Precision scores. Except Grok 3, humans generally outperform LLMs in the precision of analogical transfer. In the no hint condition, the disparity is particularly pronounced: most LLMs, whether reasoning or nonreasoning, exhibit near-zero precision. By contrast, human subjects consistently achieve precision levels of 40% or higher. The provision of a hint, however, produces a dramatic improvement in LLM performance, with reasoning models displaying the most substantial gains (zero in no hint to 90% with hint). Nevertheless, even under these conditions, human precision remains superior or at par with these reasoning-based LLMs models (in some instances reaching 100% precision). These findings underscore that, although reasoning-oriented LLMs show clear progress in the matching process when appropriately cued, human subjects continue to demonstrate a more reliable capacity for precise analogical transfer, apart from isolated cases such as Grok's performance on specific analogy types (e.g., survivorship bias in the present experimental setting).

4.2.2. Recall. The analysis of recall presents a sharply contrasting pattern to that observed for precision. In the no-hint condition, recall rates are near zero for both humans and LLMs, with the sole exception of Gemma, which exhibits perfect recall (100%). Once a hint is provided, however, the dynamic shifts dramatically: LLMs, both reasoning or nonreasoning, consistently and comprehensively surpass human subjects, achieving recall rates of 100%. This finding suggests that in a simple analogy retrieval—as distinct from precise schema matching—LLMs currently outperform humans by a significant margin.[5] However, such high capacity for retrieval also indicates a tendency toward the over-application of analogies. Such patterns may be symptomatic of a demand effect inherent in LLM behavior (Gui and Toubia 2023): once cued, these models tend to over-interpret the prompt as a directive to produce an analogy, even when a retrieved source is not necessarily the most appropriate one. This interpretive bias elevates recall but arguably risks reducing overall fidelity in analogical transfer, as it inflates retrieval without commensurate gains in matching accuracy. In fact, giving an LLM a lattice of business models, per Charlie Munger's injunction to humans, might be dangerous in the sense of eliciting many ill-fitting analogies.

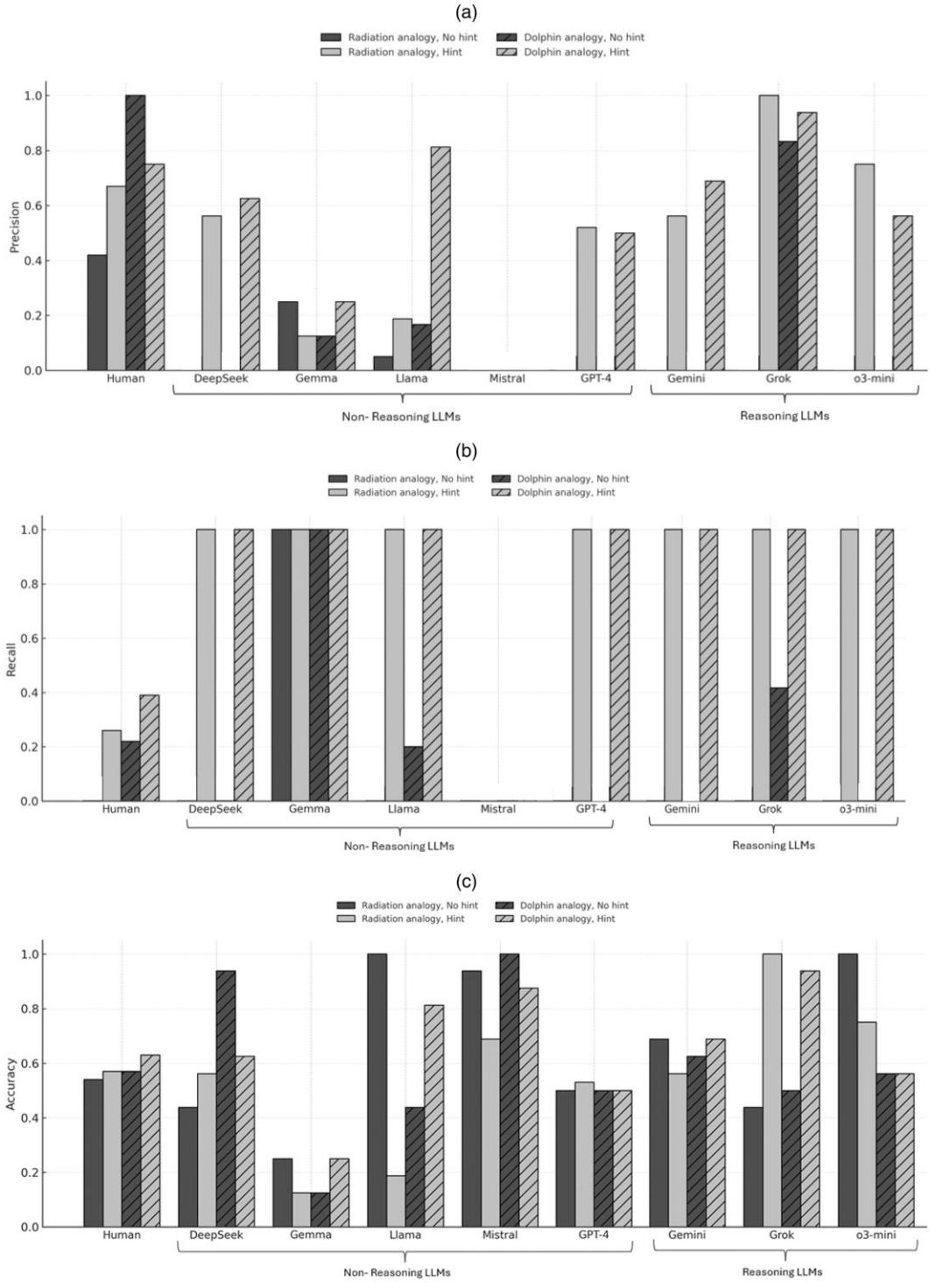
4.2.3. Matching Accuracy. In the case of overall matching accuracy, the role of the hint is far less consequential than in the improvement of precision or recall, both for humans and for most LLMs. Human participants achieve accuracies on the order of 60%, a level that surpasses many nonreasoning models, which typically fall within the 20%–50% range. A particularly striking case is Mistral: Although it achieves 80%–100% overall matching accuracy, it shows zero precision and zero recall. This apparently high matching accuracy is a statistical artefact: Mistral neither retrieves nor matches the correct analogies, yet it appears "accurate" in the trivial sense of correctly classifying negative cases, that is, rejecting nonanalogical mappings. Finally, in contrast to precision, where reasoning-oriented models consistently outperform nonreasoning ones, overall matching accuracy does not reveal a strong distinction between the two classes of LLMs.[6]

Taken together, these findings indicate that, although overall matching accuracy levels are broadly comparable across humans, nonreasoning LLMs, and reasoning LLMs, the latter occupy a particularly advantageous position. Reasoning models exhibit a balance characterized by both relatively high recall and markedly improved precision when cued, situating them at a "sweet spot" in analogical transfer performance. This balance is critical, as it demonstrates that reasoning LLMs are not only capable of retrieving candidate analogies at scale but, when appropriately guided, can also match the correct causal schemas with increasing reliability, bringing their performance closer to human-like analogical reasoning. Yet, a sharp distinction emerges in the mechanisms underpinning matching accuracy. Human performance is driven primarily by high precision—a feature that reasoning LLMs are beginning to approximate, whereas LLM performance is disproportionately anchored in high recall, which also raises the concern of a "demand effect" in LLMs, that is, a tendency to over-apply analogies.

4.3. Analogical Transfer Failure: Surface vs. Structural Similarity

In our final analysis, we examine in detail the mechanisms underlying incorrect problem solutions that arise from the misapplication of analogies. Focusing

Figure 4. Precision, Recall, and Matching Accuracy



Notes. (a) Precision of analogical transfer. Missing bars suggest that the Precision scores for those particular conditions are zero. (b) Recall of analogical transfer. Missing bars suggest that the Recall scores for those particular conditions are zero. (c) Accuracy of analogical transfer.

on the with hint condition, we human-coded each response to identify instances of analogical misapplication. Such cases were defined as transfers of the Dolphin story to the Factory problem or of the Radiation story to the HR problem, stemming either from superficial correspondences across domains or from a structural misconstrual of the underlying causal schema. The latter was operationalized as deviations from the prespecified causal schemas when mapping from source stories to target problems. Conceptually,

this distinction between surface-based and structural misapplications parallels prior theoretical treatments of analogical failure (Holyoak and Koh 1987, Gentner and Smith 2013). For example, a structural misapplication would be a claim that the Factory manager only heard about traffic problems from drivers who were late and in fact there was no problem with supplies. A surface-based misapplication would be flagged if the solution to the Factory supply problems was a solution inspired by the creatures of the sea. Figure 5 illustrates these two mechanisms of misapplication—surface versus structural—across human subjects and LLMs for both problems.

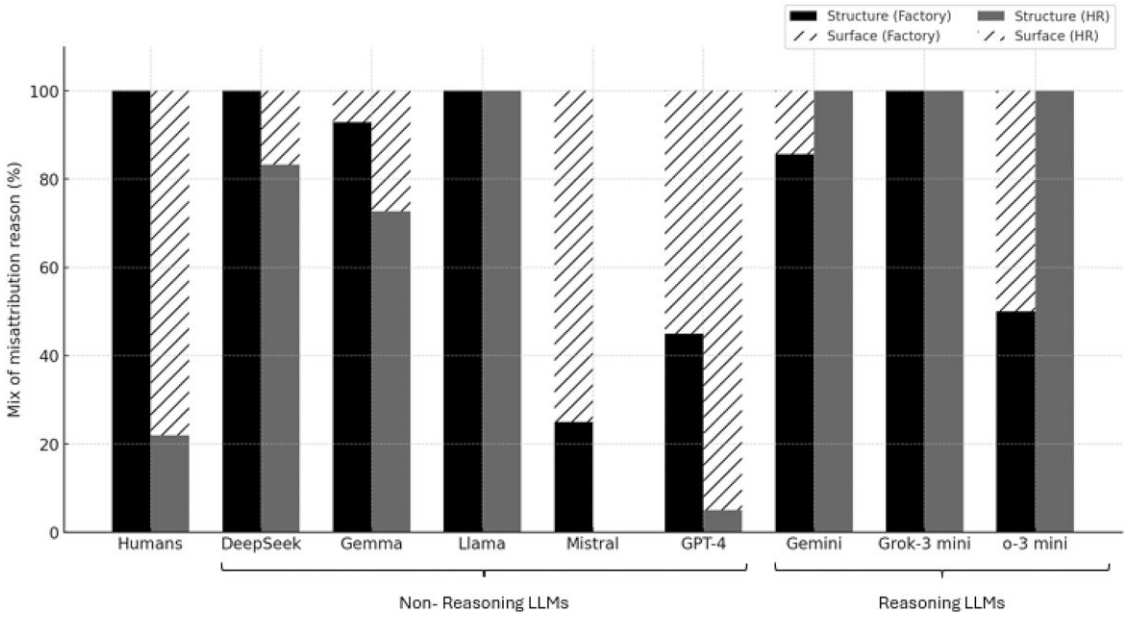
The analysis reveals fundamental differences in the mechanisms by which analogical transfer fails, distinguishing between surface-level and structural forms of mismatch. First, the type of problem strongly conditions the locus of misapplication. For both humans and nonreasoning LLMs (with the exception of Mistral), surface-level misapplication is more prevalent in the HR problem (up to 80%, compared with 10%–50% for the Factory problem). Here, responses frequently relied on superficial feature overlap rather than deeper causal alignment. For example, surface attributes such as "targeted beams" in Dr. Clarke's Radiation story were mapped onto "targeted training programs" as a proposed HR intervention. This analogy ignores the correct underlying schema of survivorship bias.

Second, reasoning-oriented LLMs display a distinctive error profile: they are substantially more likely than humans or nonreasoning models to misapply analogies through structural misconstrual of causal schemas (50%–100% of the error mix being structural mismatch; even for the HR problem). In these cases, the models appear to overfit or fabricate internally consistent causal mappings in response to the implicit "demand" to solve the problem via analogy. A recurrent example of this structural misapplication is the inference that "a one-size-fits-all approach does not work." Here, reasoning LLMs analogized from the splitting of a radiation beam into multiple smaller beams to the idea that HR training programs should not be pan-organizational but instead department specific. Although this reasoning is internally coherent, it constitutes a misalignment with the correct causal schema of survivorship bias. Examples of analogical misapplication (raw text responses that we coded) are provided in the Online Appendix (Section 5).

5. Discussion

This study set out to evaluate whether state-of-the-art LLMs can match or surpass human reasoners in analogical reasoning when tasks are framed entirely in natural language and embedded within managerial contexts. To this end, we employed a design in which participants—either business school students ($n = 199$) or eight LLM AIs ($n = 512$)—were randomly assigned to conditions in which they were presented with two candidate source analogies (radiation therapy; survivorship bias) and tasked with identifying the more structurally appropriate analogue for solving one of two novel target problems (Factory logistics; HR pilot

Figure 5. Analogy Misappropriation: Surface vs. Structure



Notes. In aggregate, 84 LLM responses were classified as instances of incorrect analogical transfer. Of these, two responses were excluded because the textual outputs from those trial runs were inconclusive for coding. Accordingly, the analysis reported here is based on 82 LLM responses across models.

study). The experimental conditions and procedures were identical across humans and LLMs; the only difference was the type of agent performing the task.

Three key patterns emerged from the data. First, our study introduced a matching problem, where participants chose between competing analogues, and this appears to have contributed to lowered task solvability relative to traditional one-to-one analogical transfer paradigms. This suggests that source matching imposes a nontrivial cognitive burden, consistent with prior theoretical treatments that distinguish retrieval and mapping from matching.

Second, when examining the conditions under which LLMs demonstrate advantages over humans, we find these benefits are rooted in analogical retrieval rather than precise matching. In conditions where the existence of an appropriate source was explicitly indicated (the hint condition), LLMs consistently achieved perfect recall, successfully identifying all cases where structurally valid sources were available. Human participants achieved substantially higher precision (consistently 40% or higher, reaching up to 100% in some conditions), meaning that when they selected a source analogy, it was typically the structurally appropriate choice. However, this came at the cost of lower recall, with humans showing reduced sensitivity to the availability of correct matches even when hints were provided. In stark contrast, LLMs displayed the reverse pattern: once appropriately cued, they achieved perfect recall but not precision (often near-zero in no-hint conditions), generating a high number of false positives in the process. Therefore, this retrieval advantage did not necessarily translate into superior overall analogical reasoning performance, as LLMs simultaneously exhibited dramatically reduced precision relative to humans (although reasoning-based models appear to be closing the gap).

Finally, the nature of these errors differed systematically across agent types, revealing distinct failure modes in analogical reasoning. Human errors typically stemmed from more conservative source endorsement and occasional misinterpretations of causal structure, suggesting a higher threshold for analogical transfer, but also susceptibility to misconstrual when structural similarities are masked by surface differences. LLMs, by contrast, exhibited a pronounced demand effect: Once cued to engage in analogical reasoning, they demonstrated a systematic tendency to over-interpret prompts as directives to produce analogical solutions, even when retrieved sources were not structurally appropriate. This interpretive bias manifested differently across LLM architectures. Nonreasoning models showed predominantly surface-level *overgeneralization*, mapping superficial features across domains without attending to deeper relational structure. Interestingly, reasoning LLMs, however, displayed a more sophisticated but equally problematic error pattern: structural misconstrual of causal schemas. These models appeared to fabricate internally consistent but incorrect causal mappings in response to the implicit demand to solve problems analogically, often constructing plausible but misaligned theoretical mapping between source and target domains.

The implications of these findings extend beyond performance metrics to reveal fundamental differences in analogical reasoning mechanisms. Although LLMs excel at analogical retrieval—identifying potential source cases at scale—their tendency toward overapplication and systematic misalignment of causal structures suggests that current architectures lack the discriminative mechanisms that allow humans to judiciously apply analogical reasoning. The emergence of reasoning-oriented models represents progress toward more precise analogical mapping, yet these advances come with their own failure modes, including the construction of sophisticated but incorrect analogical frameworks. This suggests that the path toward human-like analogical reasoning in AI systems will require not only improved retrieval and mapping capabilities, but also the development of better matching capabilities.

5.1. Implications for Analogical Reasoning in Strategic Decision Making

Although analogical reasoning has long been recognized as central to strategic cognition, most extant models emphasize retrieval and mapping, drawing heavily on insights from cognitive psychology. The third step, matching, or the selection from candidate analogues, has remained empirically underdeveloped despite its salience in theory and practice. Our findings suggest that this omission is consequential. In particular, the matching stage, where one must adjudicate among superficially plausible analogues, emerges as the locus at which human strategic reasoners retain a decisive advantage over state-of-the-art LLMs. The contrast in error profiles is telling: Human participants exhibited high precision and low recall, a pattern aligned with a "hedgehog" cognitive style (Tetlock 2017) that privileges cautious commitment and minimizes false positives (type I errors). LLMs, in contrast, adopted a more "fox-like" stance, exhibiting high recall and lower precision, thereby reducing false negatives (type II errors) at the cost of admitting spurious analogies. Further, counter to the intuition that false positives (misapplication of analogies) must arise mainly from weighing too heavily on superficial similarity when mapping between source and target, our results show that reasoning LLMs and humans are prone to seeing dubious similarities even at the structural level.

Our findings suggest that the process of selecting which candidate analogy to apply when multiple are available (i.e., matching) deserves explicit theorization alongside retrieval and mapping. Although

prior models have treated evaluation as part of mapping, our results show that the distinctive error patterns of humans and LLMs emerge precisely at this stage. Human participants displayed high precision but low recall: they rarely chose an inappropriate analogy (precision) yet often failed to recognize when a structurally valid analogy was available (recall). LLMs generally showed the opposite pattern, with high recall but low precision, with the exception of some reasoning-based models that are also high on precision: they almost never missed a relevant candidate but frequently misapplied one. These contrasting profiles highlight the essence of the matching problem: Reasoning agents must tradeoff between retrieving broadly and filtering accurately. By empirically demonstrating this precision–recall tradeoff, our study isolates matching as a decision bottleneck in analogical reasoning, distinct from mapping itself, and provides a framework for theorizing how different cognitive systems negotiate this tradeoff.

The distinct error profiles of humans and LLMs further provide microfoundational insight into the potential for human–AI complementarity in strategic decision-making contexts. A possible division of labor with specialization (Puranam 2021) may involve LLMs cast as high-throughput *generators* of plausible analogical frames, expanding the strategic option set through computational breadth. Human agents, in turn, can function as *evaluative filters*, imposing coherence by discriminating among analogues based on deeper structural alignment.

However, if human skills at retrieval are to be preserved, one might instead adopt an ensemble approach (Choudhary et al. 2025), where humans and LLMs independently retrieve and map analogies, with a comparison and selection between the final outputs of the two. By experimentally demonstrating that LLMs and humans err in systematically different ways, our findings lay the groundwork for a more fine-grained design of decision workflows in strategy that explicitly leverage this asymmetry, like allocating analogical generation to machines, while preserving final analogical selection as a human prerogative.

5.2. Implications for Managerial Practice

For humans, the act of evaluating multiple candidate analogies in parallel imposes cognitive load, which limits recall. Yet when they do engage with a candidate analogy, humans are markedly better at recognizing structural correspondence between source and target. This structural sensitivity enables them to avoid misapplying analogies even under uncertainty, yielding high precision. LLMs by contrast, no comparable cognitive cost in generating and comparing multiple analogies and thus high recall. In fact, the evidence that the more recent reasoning-based LLM models also converge with humans on the precision measure opens up interesting possibilities. Thus, although LLMs outperform humans in generating candidate analogies, humans perform with precision in evaluating causal structure, making the two systems complementary in the analogical reasoning process.

Carroll and Sørensen (2024) emphasize that analogy in strategy rests on two different bases: rhetorical vividness, often grounded in surface similarity, and problem-solving validity, which depends on detecting structural causal equivalence. Our findings map directly onto this distinction. In short, LLMs trade precision for breadth, while humans trade breadth for precision, and it is this complementarity that makes the human–AI combination potentially powerful for strategy.

The findings carry several actionable implications for the design and governance of AI-assisted strategic decision making. First, organizations deploying large language models should treat model outputs not as definitive recommendations but as *candidate sets*, that is, preliminary suggestions whose utility depends on subsequent human filtering. Interfaces that render the underlying causal schemas more transparent, such as through visualizations of structural alignments (e.g., directed acyclic graphs), can support this division of labor by combining algorithmic breadth with human depth in evaluation.

Second, the persistence of conservative analogical endorsement among human participants, driven by high thresholds for source selection, underscores the continued value of managers in exercising structural vigilance. The advent of LLMs does not reduce the need for schema-based reasoning; rather, it amplifies its significance. As machines expand the analogical search space, the human task becomes increasingly one of discerning which structural correspondences truly warrant attention. Thus, training that emphasizes systematic comparison, causal mapping, and analogical discrimination remains essential.

Third, our results suggest that managerial reliance on either humans or machines should be guided by an explicit awareness of *error cost asymmetries*. In decision contexts where the cost of missing a valid analogy outweighs the cost of entertaining false leads, such as in time-sensitive crisis response, leveraging LLMs' high recall may be desirable—although care must be taken to mitigate their tendency toward over-application when cued to engage analogically. Conversely, in high-stakes strategic decisions where false positives carry irreversible consequences, such as major capital investments or organizational pivots, human filtering, with its higher specificity, should dominate.

Finally, the findings support a governance architecture centered on *human-in-the-loop* oversight. Delegating analogical reasoning wholesale to LLMs, particularly when their outputs are driven by superficial feature

similarity, risks systematic misframing of strategic problems. Our findings suggest a complementary division of labor: LLMs can serve as powerful engines for generating candidate analogical mappings, but human managers retain a crucial role in adjudicating the validity of those mappings against the structural realities of the problem and the context. Over-reliance on reasoning LLMs may provide managers with seemingly logical causal solutions that suffer from structural misconstrual—internally coherent but incorrect mappings between source and target domains. Thus, embedding human judgment at critical evaluation junctures serves not only as a safeguard against such misalignment but also as a means of realizing the potential complementarities between machine-generated option sets and human causal reasoning.

5.3. Limitations and Avenues for Future Research

As with any exploratory study, certain limitations constrain the scope of our findings and point toward promising directions for future inquiry. First, our human participant pool consisted of business school students—a common proxy for bounded but intendedly rational decision makers in strategic cognition research, but nonetheless a sample of strategy novices. Although appropriate for isolating cognitive mechanisms, replication with senior executives or cross-cultural cohorts would enhance the external validity of our claims. The task environment itself was deliberately simplified: Participants were asked to select between two analogical sources for each of two target problems. Although this structure enabled tight control over the matching challenge, it does not reflect the complexity of real-world strategic decision making, where the analogical search space may contain dozens of plausible comparisons. Scaling the number of candidate sources and systematically varying the search costs could illuminate how human and AI capabilities interact under conditions of greater ecological realism.

Second, our study employs a discrete "correct/incorrect" coding of analogical transfer, which may appear restrictive relative to the open-ended, appropriateness-based evaluation of analogies often emphasized in strategic contexts. This was a deliberate assumption made for tractability within our two-source, two-problem design, which required us to specify in advance which source–target pairings preserved structural relations, and which did not. Framing the task this way allowed us to create a controlled setting in which human and LLM performance could be compared along commensurable dimensions. In real strategic settings analogies exist on a continuum of appropriateness rather than a binary scale, and our results should be interpreted with that scope condition in mind.

Third, as with most of the studies that explore the capabilities and applications of state-of-the-art LLMs, our findings are necessarily time bound. What constitutes state-of-the-art continues to evolve rapidly. We observed that with the advent of reasoning AI models, the quality and characteristics of the answers have changed and often improved. This study thus has to be seen as a snapshot of a rapidly evolving technological landscape. The specific performance gaps and complementarities we identify between human and AI capabilities may shift as models advance, potentially altering the relative advantages we document. Future research should revisit these comparisons periodically to track how evolving AI models reshape the division of cognitive labor in strategic decision making. Rather than offering definitive conclusions about AI's role in strategy, our work establishes a methodological foundation and baseline for ongoing assessment as both AI systems, and their organizational applications mature.

Finally, although we employed causal schema representations, intercoder reliability checks to evaluate the quality of analogical transfers for humans, and a replicable approach to detect causa schema use in the LLM responses, the assessment of structural fit still involved a degree of qualitative judgment. Incorporating automated schema-matching techniques or embedding model-generated rationales into the evaluation process could enhance reproducibility in future datasets. Collectively, these extensions would not only assess the robustness of the current findings but also help refine theoretical and practical understandings of human–AI complementarity in analogical reasoning.

Considering future research directions, a natural next step would be to examine how human subjects use or misuse the analogical candidates surfaced by LLMs. Although our study compared the quality of human-generated and AI-generated analogies, it did not investigate whether decision makers effectively integrate AI-generated analogies into their reasoning processes or whether they fall prey to automation bias, surface-level similarities, or other cognitive traps when presented with machine-generated options. Furthermore, research could explore how the presentation and framing of AI output influences analogical reasoning quality. For instance, do structured formats, such as highlighting structural alignments, providing multiple contrasting analogies, or explicitly flagging potential mismatches, help managers evaluate and apply analogies more critically? Conversely, what design choices might inadvertently amplify biases or reduce cognitive engagement? Understanding these human–AI interaction dynamics would provide practical guidance for developing decision support systems that enhance rather than impair strategic reasoning.

6. Conclusion

Notwithstanding the limitations outlined above, this study offers three contributions to the literature on strategic cognition. First, it foregrounds *matching*—the evaluative act of selecting structurally appropriate analogies—as a central and distinct component of analogical reasoning in strategic decision making. Although prior work has emphasized retrieval and mapping, our findings suggest that it is in the evaluative phase where reasoning quality is most meaningfully differentiated, and where human expertise retains an enduring comparative advantage. Second, the study provides the first direct empirical evidence that large language models and human strategists exhibit *complementary but asymmetric* analogical capabilities: LLMs expand the analogical search space through high recall, whereas humans contribute discriminative value through high precision. This asymmetry is not a liability but a design opportunity.

Third, we translate this insight into practical implications for the architecture of AI-assisted decision-making processes. Rather than replacing human strategic reasoning, LLMs may serve as high-throughput generators of candidate analogies, with human agents functioning as evaluative filters to preserve causal coherence. As language models continue to evolve in scale and fluency, the central challenge will not be whether they can reason analogically, but under what conditions their broad but noisy suggestions can be productively integrated with the narrower but more causally grounded judgments of human decision makers. Designing decision workflows that systematically leverage this complementarity constitutes a critical frontier for both future research and managerial practice.

Acknowledgments

The authors thank Professor Todd Zenger and three anonymous reviewers for constructive feedback throughout the review process; the ESSEC Research Centre (CERESSEC) for generous support; Maren Mickeler and Pooyan Khashabi for help with experimental data collection; and Sai Mihir Jakkaraju and Shubham Kumar for excellent research assistance. P. Sen and P. Puranam dedicate this paper to Dr. Nikhil Madan. His sudden passing has meant the loss of a brilliant mind and a genuine friend. His legacy as a researcher and thinker continues to inspire us.

Endnotes

[1] A model, even a formal model, can be understood as an analogy in a broader sense: It is a structured mapping from one causal network (the abstract world described by the model) onto another (the focal domain where the model is applied). What distinguishes models from simpler analogies is their codification and complexity (Lave and March 1975, Gilboa et al. 2014, Knudsen et al. 2019). Models are thus a subset of analogies that rise to a higher degree of specification and systematization, whereas simpler analogies may remain informal or rhetorical.

[2] For the ease and plausibility scores, check the Online Appendix (Section 3). Randomization checks confirmed that experimental groups were well balanced. No significant differences emerged between conditions for gender distribution ($\chi^2 = 6.51$, $p = 0.688$), story comprehension ratings (all $F < 0.88$, $p > 0.45$, $\eta^2 < 0.014$), or story plausibility ratings (all $F < 2.00$, $p > 0.12$, $\eta^2 < 0.03$), indicating successful randomization.

[3] An alternative approach would have been to code responses manually, providing a subjective, human-judged assessment of whether the correct analogy was applied (Gick and Holyoak 1980, 1983). Indeed, in an earlier version of the analysis—focused only on human responses and GPT-4—we adopted this method. However, with the present inter-LLM study comprising 512 responses, such coding would be both prohibitively cumbersome and vulnerable to coder subjectivity. We therefore employ the automated, objective measure described above using GPT-5. Importantly, as a robustness check against our earlier human-coded analysis, we applied GPT-5's semantic-similarity algorithm and observed very high convergence: 82% for human subjects and up to 93% for LLMs. Details of this robustness exercise are provided in the Online Appendix (Section 2.1).

[4] In this analysis, we treat the MBA responses as realizations from "Human" distribution.

[5] An outlier in this pattern is Mistral, which records zero performance on both precision and recall, underscoring the need to test this LLM under other experimental conditions.

[6] Detailed Precision, Recall, and Matching Accuracy scores of all LLM models and human subjects under the different conditions are captured in the Online Appendix (Section 4).

References

Battilana J, Obloj T, Pache AC, Sengul M (2022) Beyond shareholder value maximization: Accounting for financial/social trade-offs in dual-purpose companies. *Acad. Management Rev.* 47(2):237–258.

Blanchette I, Dunbar K (2001) Analogy use in naturalistic settings: The influence of audience, emotion, and goals. *Memory Cognition* 29(5):730–735.

Brandenburger A (2017) Where do great strategies really come from? *Strategy Sci.* 2(4):220–225.

Camposampiero G, Hersche M, Wattenhofer R, Sebastian A, Rahimi A (2025) Can large reasoning models do analogical reasoning under perceptual uncertainty? Preprint, submitted March 14, https://arxiv.org/abs/2503.11207.

Carroll GR, Sørensen JB (2024) Strategy theory using analogy: Rationale, tools and examples. *Strategy Sci.* 9(4):483–498.

Choi J, Levinthal D (2023) Wisdom in the wild: Generalization and adaptive dynamics. *Organ. Sci.* 34(3):1073–1089.

Choudhary V, Marchetti A, Shrestha YR, Puranam P (2025) Human-AI ensembles: When can they work? *J. Management* 51(2):536–569.

Cornelissen JP, Holt R, Zundel M (2011) The role of analogy and metaphor in the framing and legitimization of strategic change. *Organ. Stud.* 32(12):1701–1716.

Correa JD, Lee S, Bareinboim E (2022) Counterfactual transportability: A formal approach. *Proc. Internat. Conf. Machine Learn.* (PMLR, New York), 4370–4390.

Csaszar FA (2013) An efficient frontier in organization design: Organizational structure as a determinant of exploration and exploitation. *Organ. Sci.* 24(4):1083–1101.

Cummins DD (1992) Role of analogical reasoning in the induction of problem categories. *J. Experiment. Psych. Learn. Memory Cognition* 18(5):1103.

DeepSeek-AI (2024) DeepSeek-V3 technical report. Preprint, submitted December 27, https://arxiv.org/abs/2412.19437.

Denrell J (2003) Vicarious learning, undersampling of failure, and the myths of management. *Organ. Sci.* 14(3):227–243.

Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Ganapathy R (2024) The Llama 3 herd of models. Preprint, submitted July 31, https://arxiv.org/abs/2407.21783.

Duhaime IM, Schwenk CR (1985) Conjectures on cognitive simplification in acquisition and divestment decision making. *Acad. Management Rev.* 10(2):287–295.

Forbus KD, Gentner D, Law K (1995) MAC/FAC: A model of similarity-based retrieval. *Cognitive Sci.* 19(2):141–205.

Gary MS, Wood RE, Pillinger T (2012) Enhancing mental models, analogical transfer, and performance in strategic decision making. *Strategic Management J.* 33(11):1229–1246.

Gavetti G, Rivkin JW (2005) How strategists really think. *Harvard Bus. Rev.* 83(4):54–63.

Gavetti G, Rivkin JW (2007) On the origin of strategy: Action and cognition over time. *Organ. Sci.* 18(3):420–439.

Gavetti G, Levinthal DA, Rivkin JW (2005) Strategy making in novel and complex worlds: The power of analogy. *Strategic Management J.* 26(8):691–712.

Gemma Team (2025) Gemma 3 Technical Report. Preprint, submitted March 25, https://arxiv.org/abs/2503.19786.

Gentner D (1983) Structure-mapping: A theoretical framework for analogy. *Cognitive Sci.* 7(2):155–170.

Gentner D, Smith LA (2013) Analogical learning and reasoning. *The Oxford Handbook of Cognitive Psychology* (Oxford University Press, Oxford, UK).

Gentner D, Holyoak KJ, Kokinov BN, eds. (2001) *The Analogical Mind: Perspectives from Cognitive Science* (MIT Press).

Gentner D, Rattermann MJ, Forbus KD (1993) The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psych.* 25(4):524–575.

Gick ML, Holyoak KJ (1980) Analogical problem solving. *Cognitive Psych.* 12(3):306–355.

Gick ML, Holyoak KJ (1983) Schema induction and analogical transfer. *Cognitive Psych.* 15(1):1–38.

Gilboa I, Postlewaite A, Samuelson L, Schmeidler D (2014) Economic models as analogies. *Econom. J.* 124(578):F513–F533.

Goldwater MB, Gentner D (2015) On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. *Cognition* 137:137–153.

Google DeepMind Team (2025) Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Preprint, submitted July 7, https://arxiv.org/abs/2507.06261.

Gui G, Toubia O (2023) The challenge of using LLMs to simulate human behavior: A causal inference perspective. Preprint, submitted December 24, https://arxiv.org/abs/2312.15524.

Helfat CE, Peteraf MA (2003) The dynamic resource-based view: Capability lifecycles. *Strategic Management J.* 24(10): 997–1010.

Hofstadter DR (2001) Analogy as the core of cognition. *The Analogical Mind: Perspectives from Cognition Science*, 499–538.

Holyoak KJ, Koh K (1987) Surface and structural similarity in analogical transfer. *Memory Cognition* 15(4):332–340.

Holyoak KJ, Thagard PR (1989) Analogical mapping by constraint satisfaction. *Cognitive Sci.* 13(3):295–355.

Holyoak KJ, Thagard PR (1995) A Cognitive Architecture for Solving Ill-Defined Problems (No. ARIRN9524). https://www.researchgate.net/profile/Keith-Holyoak/publication/235117010_A_Cognitive_Architecture_for_Solving_III-Defined_Problems/links/02bfe513f42693f0f6000000/A-Cognitive-Architecture-for-Solving-III-Defined-Problems.pdf.

Jehiel P (2005) Analogy-based expectation equilibrium. *J. Econom. Theory* 123(2):81–104.

Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas DDL, Sayed WE (2023) Mistral 7B. Preprint, submitted October 10, https://arxiv.org/abs/2310.06825.

Knudsen T, Levinthal DA, Puranam P (2019) A model is a model. *Strategy Sci.* 4(1):1–3.

Koçak Ö, Puranam P (2022) Separated by a common language: How the nature of code differences shapes communication success and code convergence. *Management Sci.* 68(7):5287–5310.

Lave CA, March JG (1975) *An Introduction to Models in the Social Sciences* (University Press of America, Lanham, MD).

Leiblein MJ, Reuer JJ, Zenger T (2018) What makes a decision strategic? *Strategy Sci.* 3(4):558–573.

Levinthal DA, Schliesmann D (2025) Cautious exploitation: Learning and search in problems of evaluation and discovery. *Organ. Sci.* 36(2):903–917.

Lewis M, Mitchell M (2024) Evaluating the robustness of analogical reasoning in large language models. Preprint, submitted November 21, https://arxiv.org/abs/2411.14215.

Lovett A, Forbus K (2017) Modeling visual problem solving as analogical reasoning. *Psych. Rev.* 124(1):60.

March JG (1991) How decisions happen in organizations. *Human Comput. Interactions* 6(2):95–117.

Markman AB, Gentner D (1993) Structural alignment during similarity comparisons. *Cognitive Psych.* 25(4):431–467.

Miller KD, Lin SJ (2015) Analogical reasoning for diagnosing strategic issues in dynamic and complex environments. *Strategic Management J.* 36(13):2000–2020.

Navis C, Glynn MA (2011) Legitimate distinctiveness and the entrepreneurial identity: Influence on investor judgments of new venture plausibility. *Acad. Management Rev.* 36(3): 479–499.

Olguín MV, Tavernini LM, Trench M, Minervino RA (2022) The effect of surface similarities on the retrieval of analogous daily-life events. *Memory Cognition* 50(7):1399–1413.

OpenAI (2023) GPT-4 technical report. Preprint, submitted March 15, https://arxiv.org/abs/2303.08774.

OpenAI (2025) OpenAI o3-mini System Card. *OpenAI* (January 31), https://cdn.openai.com/o3-mini-system-card-feb10.pdf.

OpenRouter (2025) OpenRouter API and models catalog. Retrieved October 29, https://openrouter.ai/models.

Pan SJ, Tsang IW, Kwok JT, Yang Q (2010) Domain adaptation via transfer component analysis. *IEEE Trans. Neural Networks* 22(2): 199–210.

Page SE (2018) *The Model Thinker* (Basic Books, New York).

Pearl J, Bareinboim E (2022) External validity: From do-calculus to transportability across populations. *Probabilistic and Causal Inference: The Works of Judea Pearl*, 451–482.

Puranam P (2021) Human–AI collaborative decision-making as an organization design problem. *J. Organ. Design* 10:75–80.

Richland LE, Begolli KN (2016) Analogy and higher order thinking: Learning mathematics as an example. *Policy Insights Behav. Brain Sci.* 3(2):160–168.

Rindova VP, Martins LL (2023) Moral imagination, the collective desirable, and strategic purpose. *Strategy Sci.* 8(2):170–181.

Santos FM, Eisenhardt KM (2009) Constructing markets and shaping boundaries: Entrepreneurial power in nascent fields. *Acad. Management J.* 52(4):643–671.

Schilling MA (2018) The cognitive foundations of visionary strategy. *Strategy Sci.* 3(1):335–342.

Simon HA (1997) *Models of Bounded Rationality: Empirically Grounded Economic Reason*, vol. 3 (MIT Press, Cambridge, MA).

Sun B, Saenko K (2016) Deep coral: Correlation alignment for deep domain adaptation. *Proc. Computer Vision–ECCV 2016 Workshops* (Springer International Publishing, Cham, Switzerland), 443–450.

Tetlock PE (2017) *Expert Political Judgment: How Good is It? How Can We Know?* New Edition (Princeton University Press, Princeton, NJ).

Tripsas M, Gavetti G (2000) Capabilities, cognition, and inertia: Evidence from digital imaging. *Strategic Management J.* 21(10–11): 1147–1161.

Tschang FT, Ertug G (2016) New blood as an elixir of youth: Effects of human capital tenure on the explorative capability of aging firms. *Organ. Sci.* 27(4):873–892.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Polosukhin I (2017) Attention is all you need. *Adv. Neural Inf. Process. Systems* 30.

Vendetti MS, Matlen BJ, Richland LE, Bunge SA (2015) Analogical reasoning in the classroom: Insights from cognitive science. *Mind Brain Ed.* 9(2):100–106.

Webb T, Holyoak KJ, Lu H (2023a) Emergent analogical reasoning in large language models. *Nature Human Behav.* 7(9): 1526–1541.

Webb T, Fu S, Bihl T, Holyoak KJ, Lu H (2023b) Zero-shot visual reasoning through probabilistic analogical mapping. *Nature Comm.* 14(1):5144.

Yasunaga M, Chen X, Li Y, Pasupat P, Leskovec J, Liang P, Zhou D (2023) Large language models as analogical reasoners. Preprint, submitted October 3, https://arxiv.org/abs/2310.01714.

Yuan S, Chen J, Ge X, Xiao Y, Yang D (2023) Beneath surface similarity: Large language models make reasonable scientific analogies after structure abduction. Preprint, submitted May 22, https://arxiv.org/abs/2305.12660.

Zollo M, Winter SG (2002) Deliberate learning and the evolution of dynamic capabilities. *Organ. Sci.* 13(3):339–351.

Prothit Sen is an assistant professor of strategy at the Indian School of Business, Hyderabad. His research focuses on corporate strategy and organization design problems that involve artificial intelligence–human collaboration. He received his PhD in management from INSEAD.

Maciej Workiewicz is an associate professor of management at ESSEC Business School. He studies behavioral strategy, organizational learning and adaptation, and the role of organizational structure in generating collective intelligence in organizations. He received his PhD from INSEAD.

Phanish Puranam is the Roland Berger chaired professor of strategy and organization design at INSEAD. His current research explores how digital algorithms are shaping organizations as tools, templates, and teammates.