Strategy Science

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

Mean Articulation Machines

Russ McBride

To cite this article:
Russ McBride (2026) Mean Articulation Machines. Strategy Science 11(1):31-54. https://doi.org/10.1287/stsc.2025.0439

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

STRATEGY SCIENCE
Vol. 11, No. 1, March 2026, pp. 31–54
ISSN 2333-2050 (print), ISSN 2333-2077 (online)

Mean Articulation Machines

Russ McBride[a]

[a] Department of the Management of Complex Systems, University of California, Merced, California 95343
Contact: russ.mcbride@ucmerced.edu, https://orcid.org/0000-0003-2808-3082 (RM)

Received: May 1, 2025
Revised: September 22, 2025; December 3, 2025; January 15, 2026
Accepted: January 17, 2026
Published Online in Articles in Advance: March 19, 2026

https://doi.org/10.1287/stsc.2025.0439

Abstract. The performance of large language models (LLMs), both good and bad, derives from their core architecture as text pattern detection and generation machines that are sensitive to the frequency of the data upon which they are trained. They are amazing "mean articulation machines" in this sense. Using conceptual analysis and recent benchmark data, the paper identifies those strategic tasks that fall within the reliable competence of LLMs and those that remain fundamentally misaligned with LLM's associationistic architecture. The result is a practical continuum identifying where LLMs offer genuine leverage and where human cognition remains indispensable. The most challenging tasks—novel scientific and strategic breakthroughs—are currently out of reach for LLMs because of inherent limitations in their architecture. Because breakthroughs are described with text does not imply that we can simply mine text for the next novel breakthrough. In clarifying the boundary of current LLM capabilities, the paper aims to help strategic decision makers deploy these tools more effectively as powerful assistants for the majority of tasks that lie on the tractable side of the continuum.

History: Accepted for the Special Issue: Can AI Do Strategy?
Supplemental Material: The online appendix is available at https://doi.org/10.1287/stsc.2025.0439.

Keywords: AI • extreme transformational creativity • large language models • combinatorial creativity • decision making • entrepreneurship • strategy formulation • strategy implementation • transformational creativity

1. Introduction

The long artificial intelligence (AI) winters of previous decades now seem like distant history, and there is renewed optimism around deep neural networks and large language models (LLMs). Enthusiasm for the prospects of AI is back in full bloom as is, apparently, an "AI spring." The quip of Nobel laureate Robert Solow (1987) that "[y]ou can see the computer age everywhere but in the productivity statistics" now seems obsolete in the face of LLMs that can write reports, perform sophisticated data analysis tasks, answer questions better than IBM's Jeopardy system, pass the Family Medicine Board examination, and pass the Multistate Performance Law Test (Bommarito and Katz 2022, Hanna et al. 2024). But, LLMs have not achieved much-hyped "artificial general intelligence" (AGI) or "strong AI."[1] Additionally, there is confusion about LLM's appropriate domain of application, including whether AI models, like LLMs, might be able to perform strategic decision making, especially the kind of decision making that could lead to the implementation "of a value creating strategy not simultaneously being implemented by any current or potential competitors" (Barney 1991).

An LLM is a tool, and like any tool, its proper deployment should be guided by an understanding of where and when to use it. But, the challenge for anyone trying to use LLMs to create or analyze a strategy is understanding its peculiar ability profile and the sometimes wildly oscillating quality of its responses, offering, for example, the most amazing, articulate summary of some strategic framework just before failing a child's simple letter-shifting game (McCoy et al. 2023).

In what follows, I will suggest that the "secret" to understanding LLM behavior—both its superpowers and its kryptonite—boils down to a simple architectural fact; they are designed to discern multitudes of otherwise indiscernible linguistic patterns through association analysis and to deploy recombinations of them biased toward high-frequency occurrences in their training data. It would, therefore, be more accurate to refer to them as, "high-frequency-biased text pattern redeployment machines," but that is a mouthful. For a simpler (and less accurate) shorthand, I will highlight this architectural feature by referring to them as *mean articulation machines*. LLMs are amazing at articulating the center of gravity of a domain of discourse circumscribed by some prompt-triggered subset of its relevant training data. Additionally, their responses are structured by language patterns that feel natural and familiar to us. Their feats here, it is suggested, are more than worthy of the hype.

Can LLMs built around this architecture "do" strategy? In many ways, yes. Strategic decision making is of

course not a unified single skill but requires a wide variety of cognitive abilities (Ansoff 1965, Mintzberg and Waters 1985, Shoemaker 1996, Powell et al. 2011, Gavetti 2012). A continuum of increasingly difficult strategic cognition skills will be explored, which will illustrate how far along that continuum the LLM "abilities" extend. Existing benchmarks since 2021 were built to test competency across broad domains, like language understanding (SuperGlue) or graduate-level science and engineering (GPQA), and there are some benchmarks appropriate for strategy tasks as we will see. But, many of them are not specific enough to provide guidance about the deployment of an LLM for a task by some end user, like a manager or strategic decision maker. The purpose here is to offer a little guidance, at least until more strategy- and management-specific benchmarks are developed.

LLMs have a superpower—the dexterous mastery of hidden linguistic pattern that even linguists have yet to fully discern. The direct implication of this superpower is that to the degree that some strategic task depends upon well-established knowledge from widely available information, then they excel. But, to the degree that some cognitive task is orthogonal to or cannot be approached through these text association patterns and depends upon skills that lie beyond the core strength of LLMs further along the continuum of task difficulty, then one should adjust one's expectations for a successful outcome. In what follows, I will attempt to show why mean articulation is such an incredible achievement, sketch a continuum of task difficulty, and identify the fuzzy boundary beyond which LLMs cannot cross. Understanding not just their task abilities but why they have such abilities (and disabilities) should allow a strategist to predict how well an LLM will perform on almost any task and to enable more productive use of that LLM. In the end, we will see that on the most challenging of tasks—strategic and scientific breakthroughs—their superpower is a superweakness.

2. The Power of Implicit Text Patterns

LLMs derive their power from their amazing mastery of implicit patterns discoverable in massive corpora of text, including the entirety of the accessible internet. Modern LLMs are built using a neural network (McCulloch and Pitts 1943), which is a system loosely inspired by neural connections in the brain and trained to find patterns, typically now patterns in language. It was improved upon with the Rumelhart et al. (1986) gradient descent training technique, which remains the primary method today for updating parameters. The "deep learning" term of modern LLMs refers to the fact a neural network includes many, sometime hundreds, of hidden layers between the input layer and the output layer and now, often close to a trillion weighted connections (parameters). At its core, a neural network is just a very large collection of math functions stacked in layers, where each function adjusts slightly based on what it learns. At the core, LLMs are trained to predict the next most probable token—roughly a word or subword unit—in a sequence based on preceding context using an design known as the Transformer architecture (and multihead self-attention mechanism), both announced in a breakthrough paper by Vaswani et al. (2017). This objective may seem simple, but when scaled, it enables the generation of syntactically correct, semantically coherent, and contextually plausible language across a range of domains in the training data.

But, the astonishing advances in recent performance were critically the result of throwing huge amounts of data and compute power at the simple linear algebra that comprises much of an LLM. Quoting the chief exective officer (CEO) of DataBricks Ali Ghodsi: "The algorithms we're using now, they're from the '80s and '70s. But if you have enough data and enough compute power, they become magical" (Cai 2023). The math at the heart of LLM is astonishingly simple, so simple as to feel anticlimatic and depressing to some lifelong AI researchers like Sutton (2019), who famously talked about the "bitter lesson" from this, and Hofstadter (2023), or Togelius and Yannakakis (2024), who suggested "survival strategies for depressed AI academics" in the face of a depressingly simple and anticlimactic architecture.

I will focus on the core LLM activity—the dexterous discovery and recreation of hidden linguistic patterns in vast corpora of text that even linguists have yet to fully discern.[2] Any basic database is capable of standard text retrieval, which is the simplest, most trivial function that an LLM can (usually) also perform. One amazing aspect of LLMs is the ability to treat syntactically different text sequences as if they were semantically equivalent. "The dog ate the steak" is syntactically different from "the steak was eaten by the dog" and is different from "among the things the dog ate there was a steak." Massive text analysis revealed to the LLMs that these phrasings are treated as roughly identical, and so, LLMs do the same. Although it might seem trivial, treating these (and other much more complex variants) as conceptually identical was an incredible advance. Without this, any interpolation of the training data in ways relevant to human conceptual structures could not get off the ground.[3]

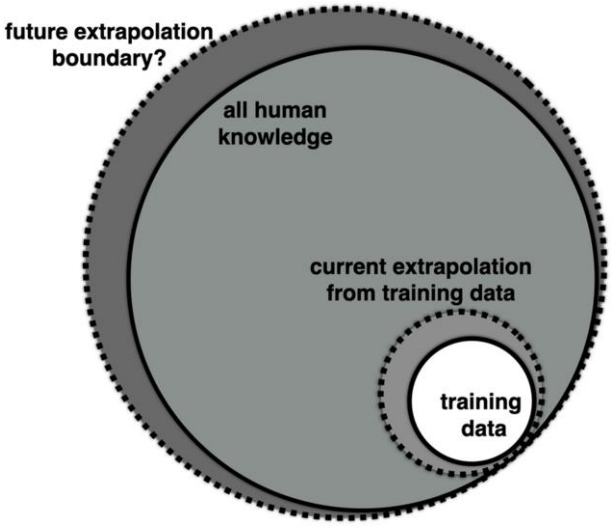
This accomplishment was made possible because an LLM captures and can reproduce the distribution tendencies of human discourse across wide ranges of granularity (word, sentence, paragraph, narrative arc, theories, ideas, styles, structure, etc.)—not because it understands language or literally knows that such expression mean anything. The patterns discovered by the model emerge from the frequency, co-occurrence of terms that often appear together, and contextual

structures (e.g., statistical dependencies beyond the immediate sentence) of the language data that it is trained on. The result is a model that learns complex patterns of linguistic co-occurrence across billions of documents (Brown et al. 2020).

It is difficult to overstate the importance of this or the extent of its implications. Google rose to search engine dominance with the simple insight that web page backlinks can serve as a proxy measure for the relevance of search results. The powerful technical trick here is *the use of implicit patterns in the enormous dispersement of text symbols to serve as a proxy for the structure of human knowledge because human knowledge is expressed through such patterns*. The unspoken ambition is as follows. If human knowledge leads inevitably to further advances in human knowledge, so too should the proxy structures of human knowledge (these implicit text patterns) lead to ever-expanding capture of and advances in human knowledge. The interesting, live question for machine learning engineers then is this: "How far can LLM models extrapolate beyond the text and text patterns upon which they are trained?" (See Figure 1.)

Optimism drives grand ambitions that LLMs (or some kind of machine learning system) can first, accurately capture all of the knowledge stored in the vast corpora of documented knowledge; second, extend to all current *undocumented* human knowledge; and third, perhaps lead to advances in science, humanities, and civilization-altering genuine knowledge breakthroughs. How far could such patterns carry us? All of the way to new scientific discoveries, novel business strategies, and new organizing principles for societies? If we think of a ship passing far from a coastline as a difficult-to-reach piece of yet unknown knowledge, then the question is this: "How far away can ripples from the wake of that ship (as the implicit patterns it gives off) be used to predict its existence?" Are the ripples in existing text enough to reflect the as-yet undiscovered knowledge that it hides between the lines? Undocumented knowledge? Future breakthroughs?

Figure 1. Training Data, Human Knowledge, and Extrapolation Boundaries



The ultimate extent to which an LLM can extrapolate beyond its training data is currently unknown and heavily debated. LLM optimists (e.g., Kaplan et al. 2020, Kurzweil 2022, Bubeck et al. 2023, OpenAI 2023, Uehara 2025) argue that the range will in the future extend far beyond the boundaries of current human knowledge and into radical advances not yet known (the larger dotted circle in Figure 1). Those with more conservative views, like Wooldridge (2020), Bender et al. (2021), LeCun (2022a, b), and Marcus (2022), suggest that current *interpolation* does not even extend to proper handling of the existing training data, much less the current boundary of existing human knowledge. A middle ground view suggests that there is a modest degree of extrapolation beyond the training data (represented with the smaller dotted circle in Figure 1).

Expanding the size of an LLM's extrapolation boundary is the ongoing goal of LLM engineers. Live research questions are as follows. "Where exactly is the boundary of what can currently be extrapolated by LLMs?" "Can that boundary extend to genuine knowledge breakthroughs?" "Can they extrapolate beyond their training data into undocumented knowledge?" "How far can they extrapolate beyond their training data?" Part of the hope of this manuscript is to sketch a plausible outline of an answer to the first question and perhaps, shed some light on the others.

3. "Mean" Articulation Machines

It seems to many as if the long unbridged chasm between syntax and semantics, most clearly described in the Searle (1980, 1990) "Chinese room argument," has finally been crossed as shown by "the dog ate the steak" example. Or has it? LLMs are statistical mirrors of human expression. The generative power of LLMs lies not in any genuine understanding of concepts but in their ability to estimate an appropriate human answer in a way that has meaning for the user. This is because it reflects the implicit, subtle patterns found in the language data upon which it was trained—all of the ones that your grammar teacher taught you and then thousands more. LLMs favor responses that reflect what has most frequently been said or written in the training data. The training process itself reinforces this bias toward centrality. High-frequency patterns are learned quickly and heavily weighted in the model's parameters (McCoy et al. 2023). Rare associations or outlier perspectives—whether innovative scientific ideas, countercultural arguments, or niche linguistic usages—are underrepresented and less influential in

shaping model behavior (Bender et al. 2021, Marcus 2022, McCoy et al. 2023, Bridgelall 2024, Zhang et al. 2024). The result is a machine that contains the full distribution of the training data, but it is "norm conforming" and heavily biased toward high-frequency ideas.

LLMs need some way of culling the infinite number of potential responses down to some manageable set. This occurs during a forward pass through the neural network after a triggering prompt, which produces a probability distribution over potential continuation words (tokens). Bridgelall (2024) observed that token frequency distributions in training data follow a power-law pattern; a very small number of tokens dominate the probability mass, whereas the majority of possible tokens fall off rapidly into the long tail of near-zero probability. In practice, this means that the model overwhelmingly favors a restricted subset of high-frequency continuations and largely ignores the vast majority of alternatives. Zhang et al. (2024) similarly confirmed that LLMs assign higher probability to text tokens that appear more often in the training text. McCoy et al. (2023) showed that LLMs are more likely to succeed on high-frequency patterns and fail to answer questions or solve problems and on less frequently seen ones.

Next, it is then the job of the "decoder" to choose the output from this restricted set of possible outputs using a statistical measure like "top-p" (nucleus) sampling, perhaps combined with "top-k" sampling or temperature scaling (Holtzman et al. 2020). Although these statistical sampling strategies can introduce variation, they still operate within the high-probability region fed to them and rarely venture into the long tail. The result is that truly novel outputs—those relying on rare or atypical combinations of words—are effectively suppressed or eliminated from the model's production. This has two implications. First, it reinforces the tendency of LLMs to act as "center of gravity" articulators rather than generators of new knowledge because high-frequency formulations dominate the high-probability assignments. Second, it structurally biases the model away from outputs that would challenge existing patterns in the training corpus. Even if such patterns are logically possible or semantically coherent, their low frequency ensures that the model's forward pass probability distribution renders them practically inaccessible. Thus, the power-law distribution of probability assignments does not merely reflect conservatism in model outputs; it actively constrains the epistemic range of the model, filtering out precisely the kinds of rare or unexpected formulations that are often the seedbed of creativity and scientific novelty (Bender et al. 2021, Marcus 2022, McCoy et al. 2023, Bridgelall 2024, Zhang et al. 2024).

LLMs were really built to solve the problem of *relevance* such that responses are relevant and appropriate to end users, and they have by and large solved it.[4] Human conversation flows naturally when each step is a relevant and appropriate continuation of what preceded it, and what often makes it feel "relevant and appropriate" is whether it has been frequently mentioned. This same process in LLMs is, therefore, a useful technique that mimics (much of the time anyway) human conversation. Human conversation has a frequency "bias" toward the most commonly discussed topics within a domain of conversation, and LLMs reflect this. This can cause problems for LLMs (and humans) when the most frequently mentioned ideas happen to be false, like, for example, that ostriches stick their heads in the sand when scared. This is one of the reasons why models also receive human training. Reinforcement learning from human feedback (RLHF) is a way of training AI by having humans judge its answers and rewarding the model when it responds in ways that humans find helpful, polite, and appropriate. It makes for more natural conversation and helps to mitigate the false but frequent preference problem. RLHF is an important step in the LLM training process that greatly improves them and significantly reduces false statements.

We do not yet have a technical term to describe the huge number of vector and linear algebra calculations in the neural net that strongly favor results more often seen in the training data. I will continued to use the term "mean" in a loose metaphorical sense, even though LLMs are clearly not outputting the simple mean of the training data distribution. LLMs, however, are "mean-articulating machines" in just this sense; they are amazing at articulating the center of gravity of those data seen most frequently during training, relative to a given prompt, and expressing that center of gravity through refined patterns of linguistic expression discovered through an extensive training phase. This also explains why the epistemic range of LLMs is largely bound by the range of the training data and hence, why AI companies have a very large line item in their budgets reserved for never-ending data acquisition.

Their "associationistic" core explains why LLMs are so effective at tasks like summarization, explanation, and restatement of widely known concepts. It is this superpower of LLMs, which is simultaneously its kryptonite, that explains the hallucinations and the strange failures that we see, like the inability to solve a problem when reworded in a unfamiliar way. Wooldridge (2018), discussing his book, said:

> There's a huge body of work looking at whether [LLMs] can actually solve problems that are not just variations of something they've already seen in their training data … Is [the LLM] really originally solving a problem vs. just doing pattern recognition. At the moment, that's one of the big questions and the jury is very much out on that, and the weight of evidence

> at the moment is they are not doing problem solving. They are doing something which is much more like pattern recognition. (Bi 2025)

Mitchell (2021) has argued that this leads to systems that are "impressive mimics." They have mastered the identification and deployment of huge numbers of linguistic patterns within their training data, which enable them to manifest incredible search and articulation skills. In fact, one way of setting LLM behavior expectations more appropriately might be to think of them as the next generation of superpowered search engines rather than intelligent systems. They are adroitly articulating associations between text discovered through subtle patterns of linguistics learned through huge amounts of digitized text—much more interpolation than extrapolation.

LLMs represent the pinnacle of performance in pattern discovery in vast quantities of text. The ability to query seemingly anything from the vast expanse of documented human knowledge and get a reply in natural language represents, at least, the single greatest educational tool and knowledge resource in human history. But, there are two fundamental challenges that come along with the mean articulation architecture. It is not clear how far interpreting every problem as a text pattern association problem can get us. Can math problems be solved as text association problems? Can genuinely novel strategy ideas be imagined? As we will see below, the answers are "often yes" and "no."

This concludes the description of LLM architecture, but there is a list of historical milestones in Online Appendix A and some suggested learning material in Online Appendix B. Next, we will start to apply what we now know about LLM architecture to the construction of a strategic cognition task continuum.

4. The Task Continuum—Easiest

4.1. Search/Verbatim Retrieval, Different Expressions of a Single Idea, Aggregation, High-Frequency Consensus Ideas, Document Templates, Linguistic/Pragmatic Styles, and Strategic Framework Assessment

We examined the basic architecture of LLMs to see, specifically, how and why LLMs respond the way that they do toward the goal of understanding their performance in strategy-specific tasks. I will not pretend to offer a comprehensive list of such tasks but instead, include some of the most frequently mentioned tasks in the strategy literature historically and the subskills often needed for them. In what follows, a continuum of task difficulty will be constructed that allows for a more detailed assessment of LLM performance and offers a rough sketch of an answer to the question about "how far can LLMs extrapolate beyond their training data?" An unsurprising theme will emerge; those tasks most approachable to association-based patterns are always easier for an LLM. Those that are not are not.

We already saw that LLMs can search (in the colloquial sense) and retrieve text and verbatim quotes from training data with proper prompting (Carlini et al. 2021), and they can individuate a concept from syntactically disparate expressions of it.

4.2. Search

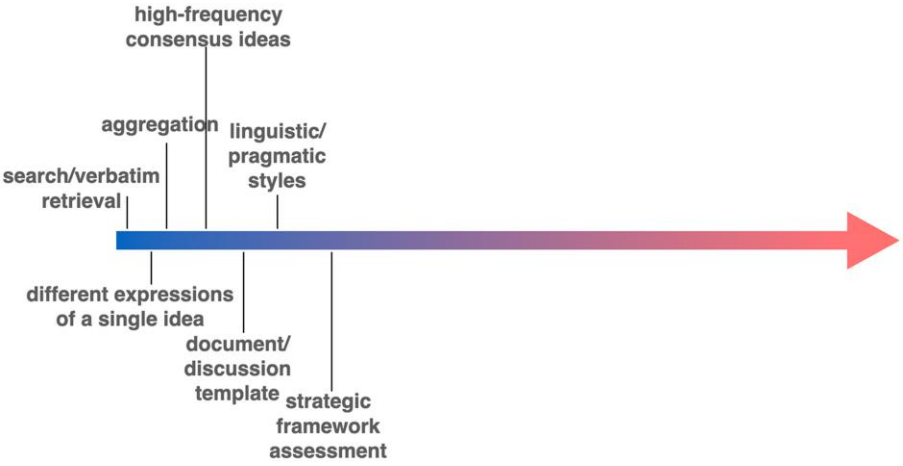
Search, in strategy, is the process of generating a broad range of strategic alternatives, solutions, or options for the firm (Cyert and March 1963, Ansoff 1965, Nelson and Winter 1982, March 1991, Levinthal 1997, Greve 2003). For example, a management team might brainstorm multiple new product ideas and market entry strategies, a task now accelerated by AI tools that can instantly draft diverse business scenarios (Csaszar et al. 2024). Given our understanding of LLM architecture, the caveat is that these business scenarios will be chosen from ones that the LLM has been exposed to in training data or relatively simple recombinations thereof. There are, it should be noted, at least three understandings of "search": standard "search" as computational find and retrieve, "search" as simple strategy ideation, and search as "novel strategy ideation" ("novel" as in "transformationally creative" (Boden 2004) as discussed below). These all must be assessed differently. The continuum of task difficulty begins with regular (colloquial) search. Simple and novel strategy ideation show up later in the continuum.

4.3. Aggregation

Aggregation refers to combining information and inputs from multiple sources or stakeholders into a collective strategic decision (Freeman 1984); for example, top managers might integrate feedback from customers, frontline employees, and regional divisions, potentially even using an AI "virtual crowd" to simulate additional stakeholder opinions in order to arrive at a well-rounded decision on a new policy (Csaszar et al. 2024). Provided that the sources are documented, this is a task where LLMs can really deploy their superpower for excellent results and offer a wide variety of aggregation results and types.

We have also noted that well-established (i.e., high-frequency) ideas are more likely to become outputs relative to outlier ideas (Mitchell 2021, Bridgelall 2024, Zhang et al. 2024). These tasks fall well within the capabilities of LLMs as does formatting a response to a document template (e.g., introduction, methods, results, and conclusion) or imitating the writing style of Wordsworth or Hemingway—a straightforward deployment of discovered text patterns built from plenty of training examples. Figure 2 shows these easy tasks on the left side of an "LLM scale of task difficulty."

Figure 2. (Color online) Easy Tasks for LLMs



An LLM can also deploy "pragmatic styles"—conventions for how language is used in interaction. This is seen in the polite disclaimers, hedges, clarifications, and refusals, reflecting conversational norms rather than raw text continuation (Bai et al. 2022, Ouyang et al. 2022). This often requires more reinforcement training from actual human feedback than other skills and makes models better at maintaining coherence, reducing repetition, and presenting arguments in stepwise fashion (aided by techniques such as chain-of-thought (CoT) prompting and unlikelihood training) (Welleck et al. 2019, Holtzman et al. 2020, Wei et al. 2022).

4.4. Strategic Framework Assessment

This involves applying frameworks to assess a firm's competitive position within the industry (Porter 1980, 2008). These are diagnostic tools that help identify the economic drivers of sustainable profitability versus mere operational effectiveness (Schendel and Hofer 1979, Rumelt 1991, Porter 1996). For instance, a strategist might employ the VRIO (valuable, rare, inimitable, organized) framework to verify if specific assets function as isolating mechanisms that block imitation rather than simply cataloging resources (Wernerfelt 1984, Dierickx and Cool 1989, Barney 1991, Peteraf 1993). Ultimately, this assessment must account for dynamic interactions and feedback loops to identify leverage points where intervention can alter the competitive landscape (Teece et al. 1997, Eisenhardt and Martin 2000, Ghemawat 2002). Modern LLMs enable more complex, data-rich framework assessments.

Performing a strategic framework assessment is an easy task. Indeed, an LLM can provide a variety of framework assessments and then provide a meta-analysis of the assessments, and the only real human work consists of supplying documents and running some prompts.

4.5. Analogical Reasoning, Combinatorial Creativity with Familiar Components, Simple Strategy Ideation, High-Frequency Information with a Conflict, Induction, and Math

We move now into more difficult tasks but are still mostly within the domain of competence of LLMs (see Figure 3).

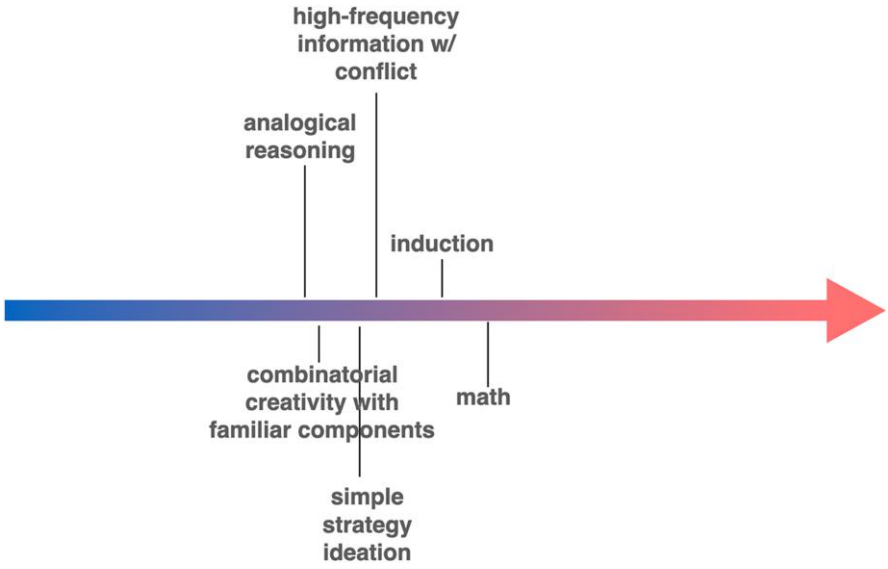
4.6. Analogical Reasoning

This involves transferring a pattern from a familiar source domain to an unfamiliar target domain (Gentner 1983), a skill seen by some as central to strategic innovation in uncertain environments (Gavetti and Rivkin 2005). For example, a fintech might reason by analogy using Apple's app ecosystem strategy as a template for how to build a developer community around the start-up's financial platform—leveraging insights from a seemingly unrelated context to solve a strategic problem at hand. (Lovallo et al. 2011, Gary et al. 2012). Reapplying an existing pattern is the bread and butter of LLM activity, and recent empirical evaluations confirm that LLMs display a remarkable proficiency in zero-shot (i.e., no examples in the prompt) analogical reasoning, frequently matching or exceeding human performance in retrieving and mapping abstract relational patterns across vast cognitive distances (Webb et al. 2023). Consequently, these models serve as one tool for overcoming the "local search" bias that often constrains human strategists to familiar but suboptimal solutions (Gavetti 2012).

4.7. Combinatorial Creativity with Familiar Components

The distinction from Boden (2004) between simple "combinatorial creativity" versus novel "transformational creativity" (to which I will add the category of "extreme transformational creativity" later) is useful here. LLMs

Figure 3. (Color online) Slightly Harder Tasks for LLMs



excel at simple combinatorial creativity, mixing ideas and patterns that they have seen, but they stumble at truly novel creativity, especially when it violates existing, established knowledge in the extreme cases (Gervás 2024, Franceschelli and Musolesi 2025). We will look at the place of each subtype—combinatorial, transformational, and extreme transformational—on our continuum of task difficulty, beginning with the simplest.

LLMs can display a limited form of creativity by recombining familiar ideas in new ways, generating outputs unlikely to have appeared verbatim in their training data. For example, they might produce the phrase "an orange-slice sandwich" by combining familiar elements—"orange slices" and "sandwiches"—using patterns of plausible composition. This is the "combinatorial creativity" of Boden (2004). LLM models often do this in fictional stories, jokes, and analogies that remix known material (Brown et al. 2020). Cognitive studies similarly find that LLMs can generate ideas that humans judge as creative (in this sense) when outputs involve rearrangements of well-represented concepts (Binz and Schulz 2023). At the same time, large-scale studies of creative writing suggest that such recombination tends to converge on common themes, producing homogeneity rather than true diversity (Wenger and Kenett 2025). LLMs remix existing knowledge into new but statistically coherent forms.

4.8. Simple Strategy Ideation

This requires combinatorial creativity and overlaps with aggregation, so it is a relatively simple task. Most LLMs with proper prompting and appropriate data can use training exposure to thousands of strategies to combine the relevant components into useful simple strategic idea. One might ask an LLM to "generate a simple strategy for increasing customer retention for a midpriced home fitness equipment company" and get reasonable suggestions, like adding a loyalty program or bundling extended warranties.

4.9. High-Frequency Ideas That Conflict with Some Information

High-frequency ideas in training data are much more common than contradictory information. When an LLM encounters debates (e.g., Copenhagen versus pilot-wave theories in physics), it reports both but gravitates toward whichever view dominates the corpus (Copenhagen in this case). When disagreement is uneven, the minority position may effectively disappear. If 90% of training data state that "Shakespeare wrote Hamlet" and 10% of training data attribute it to Marlowe, default decoding will almost always return "Shakespeare" because low-probability continuations are suppressed (Holtzman et al. 2020, Bridgelall 2024, Zhang et al. 2024). The model is not really choosing a side; it is averaging toward the statistical center.

This becomes problematic when frequency conflicts with truth. TruthfulQA (Lin et al. 2022) shows that models frequently repeat culturally common falsehoods (e.g., ostriches bury their heads when scared) because these answers are linguistically more probable than factual ones. Instruction tuning and RLHF reduce but do not eliminate this consensus bias, and later evaluations (Ji et al. 2023, OpenAI 2023) confirm that likelihood and accuracy often diverge. In short, LLMs tend to imitate widespread misconceptions whenever such misconceptions are well represented in the training data.

For tasks that merely seek a majority view, this probabilistic smoothing is usually acceptable. But, for

strategic decision making—where valuable insights often hide beneath low-frequency, contradictory, or unconventional perspectives (McBride et al. 2024)—this bias is a fundamental limitation.

4.10. Induction

Induction in strategic cognition involves generalizing from a limited set of prior observations (e.g., market episodes or competitor moves) into expectations about what will happen (in the market or with a competitor) in the future, effectively projecting a pattern from "what has happened so far" to "what is likely to happen again" (March and Simon 1958, March 1991, Gavetti 2005, Posen et al. 2018). A strategist who runs several small A/B tests on alternative pricing schemes and infers a more general rule about price elasticity is engaged in a kind of inductive inference, drawing regularities out of noisy experience to guide subsequent commitments under uncertainty (Levinthal 1997, Adner and Levinthal 2008). LLMs are, at their core, powerful statistical engines for this sort of pattern-based induction over text. Next-token training and in-context learning allow them to infer plausible rules from multiple examples and to interpolate within familiar regimes (Brown et al. 2020, von Oswald et al. 2022, Xie et al. 2022, Bai et al. 2023). Yet, their inductive competence is tightly bounded by the distributions on which they were trained; when surface cues, problem formats, or underlying mechanisms diverge from those seen in the training data, performance degrades sharply, and "induction" collapses back into frequency-driven guessing rather than principled rule formation (Binz and Schulz 2023, McCoy et al. 2023, Ren and Liu 2024). In strategy work, LLMs can, therefore, assist with inductive tasks that involve summarizing well-documented regularities (e.g., extracting common success patterns from historical cases or synthesizing empirical findings), but they remain unreliable when the focal problem sits in a genuinely novel regime where past textual patterns provide, at best, a weak and potentially misleading guide.

4.11. Math

Although mathematics is formally a deductive activity (usually), its role in strategic cognition is largely practical; strategists routinely need to compute (e.g., break-even points, contribution margins, customer lifetime value, or expected returns under different scenarios), and these calculations often structure downstream choices about pricing, marketing spend, capital budgeting, and capacity planning (Kaplan and Norton 1996, Makridakis et al. 2010). Modern LLMs handle such routine, well-specified computations effectively (though often only by delegating to python code), and recent systems have even reached gold medal-level performance on structured Olympiad-style tasks under heavy scaffolding (Cai and Singh 2025), which makes them attractive assistants for everyday quantitative strategy work.

The limitations emerge when mathematical reasoning departs from familiar templates. Even strong models are brittle on novel, proof-style, or distribution-shifted problems, and much of their high-end mathematical success appears to stem from pattern recombination and consensus-style "self-consistency" voting rather than the construction of genuinely new mathematical arguments (Wang et al. 2022, Frieder et al. 2023). This is what we would expect from machines trained to find patterns through associations. At the core, the "reasoning" is still associationistic rather than deductive. They perform well on mathematical forms that they have seen often but poorly on less frequently seen forms (even in deterministic computations), and they can be pushed into failure modes by minor changes to formatting or task frequency (e.g., linear conversion formulas that are common versus those that are equally simple but rare) (McCoy et al. 2023, pp. 13–20). This means that even simple strategic calculations can fail if the query is expressed in an unfamiliar or distribution-rare format, such as writing numbers in alternating capitals or embedding them in unusual syntactic structures.

The strategic implication is clear. LLMs are reliable assistants for routine quantitative tasks that stay within familiar representational regimes—basic profit calculations, accounting summaries, KPI roll ups—but they are less reliable when math functions as exploration, conceptual innovation, or original model building within a strategy process (Hamel and Prahalad 1994, Mintzberg 1994). Their mathematical competence is strongest where strategy needs structured interpolation and weakest where strategy requires genuine abstraction, reframing, or novel quantitative insight.

5. The Task Continuum—Harder

5.1. Abductive Reasoning, Deductive Reasoning, Causal Reasoning, Novel Idea Generation with Minimal Conflict, and Novel Strategy Ideation

Here, we move into the realm of tasks where LLMs fall far behind human performance (more than about 25% behind on most benchmarks), and when they do succeed, it is often because they have seen the problem (or a similar pattern) in their training.

5.2. Abductive Reasoning (Hypothesis Generation)

Abductive reasoning (what Sherlock Holmes actually did—not deduction) is often called "the logic of discovery," and when successful, it arrives at the most plausible explanatory hypothesis for a puzzling observation (Peirce 1955, Hanson 1958). Abduction allows

strategists to leap from incomplete or ambiguous data to a coherent explanation (Dunne and Martin 2006, Martin 2009, Sergeeva et al. 2021, Bhardwaj et al. 2025). For instance, a strategist observing an inexplicable shift in consumer preference might employ abductive inference to postulate a latent market need or a nascent competitive threat, thereby generating a testable strategic conjecture despite the absence of definitive evidence (Weick 1989, Ketokivi and Mantere 2010). This mode of reasoning is critical to strategic innovation and theory construction as it permits the generation of novel business models that cannot be derived solely from analyzing historical data (Nonaka and Takeuchi 1995, Bamberger 2018). Abduction is a well-explored subfield in old school symbolic AI, where it is called "abductive logic programming," and here again, LLMs can certainly *simulate* abduction but seem to fail when confronted with cases that do not match prior exposure. If all of the LLM training data showed that "the butler did it" and there is no butler mentioned in the case, then it will get confused. This is a controversial position, with Wei et al. (2022) and Berg et al. (2023) suggesting just the opposite.

Bhagavatula et al. (2020) introduced the first "abductive common sense reasoning" benchmark for LLMs consisting of an easier test (alphaNLI) where the model had to choose from two explanations for the presented observations and a more realistic test (alphaNLG) where an explanation for the observations had to be generated from scratch. The 2020 models performed at about a 45% level compared with 96% for humans, with current (2025) models performing much better at about 72%. Since then, more than a dozen benchmarks were developed to test adductive reasoning, and on average, the performance of the best current models is about 20%–30% behind human performance. A reasonable success boundary might be one where a model equals human performance on more than half of the benchmarks. The benchmarks with the smallest gap in human/machine performance are, perhaps unsurprisingly, tests whose pattern appeared frequently in the training data (i.e., StoryClose, HellaSwag, and SocialIQA), so these should probably be eliminated from consideration. We do not yet have a model that can best human performance here.

5.3. Deduction

Deductive reasoning matters in strategy because it applies general principles to concrete situations—for example, inferring that a specific competitor cannot sustain a price war because in general, no high-cost entrant can sustain a price war (Porter 1980) or that a resource failing a VRIO criterion cannot yield sustained advantage (Barney 1991). Strategists routinely rely on such rule-governed inference in game theory, performance diagnostics, and capability assessment (Levinthal 1997, Teece et al. 1997, Gavetti 2005).

LLMs can imitate deduction when problems resemble familiar training patterns, but they remain brittle under small rephrasings or novel instantiations. Their performance jumped, however, with the advent of chain-of-thought reasoning (Wang et al. 2022). CoT reasoning emerged in the early 2020s as researchers observed that large language models could substantially improve performance on multistep reasoning tasks when prompted to generate explicit intermediate reasoning steps (Kojima et al. 2022, Wei et al. 2022). CoT does not introduce a new model architecture; instead, it forces the model to allocate probability mass across multiple intermediate steps rather than collapsing everything into a single answer. It also guides the model to "think out loud" by showing step-by-step reasoning before giving a final answer, sometimes combined with sampling or verification methods, such as self-consistency (Wang et al. 2022). In mathematics and deductive reasoning, CoT substantially improves accuracy by enforcing stepwise constraint satisfaction and reducing omitted or inconsistent intermediate steps (Wang et al. 2022, Wei et al. 2022). By contrast, in abductive reasoning, CoT mainly improves the coherence and plausibility of explanations rather than the generation of genuinely novel or causally correct hypotheses, indicating that it enhances reasoning articulation rather than underlying inferential capacity (Kojima et al. 2022).

LLMs often miss deeper logical equivalences (e.g., contraposition) even when they can restate surface-level paraphrases (Lin et al. 2022). Classic deductive tests, such as the Wason Selection Task, confirm that models rely on surface familiarity, improving only when prompts match known templates (Lampinen et al. 2024, Seals and Shalin 2024). Chain of thought improves presentation but not underlying validity (Tang and Kejriwal 2024, Zhou et al. 2025). Apple Machine Learning Research Team (2025) describes this as an "illusion of thinking" (performance collapses once problems deviate from learned distributions).

McCoy et al. (2023) show that these failures reflect frequency sensitivity rather than rule application. They designed 11 adversarial tasks—such as encoding text with rot-13 ("rotating" each character with the letter 13 letters forward in the English alphabet), swapping the order of paragraphs, or computing linear functions—and found that performance depends not on complexity but on the training frequency of the solution format, the input, and the target output. When asked to apply a simple rule differently—such as rotating (replacing) letters by 2 letters forward in the alphabet instead of the more common 13 letters forward—LLMs succeed only if the specific variant is common in training. They failed on rot-2 and rot-8, despite the task being simple and structurally identically to the more frequently seen

rot-13 (McCoy et al. 2023, pp. 13–15). Replacing each letter with the letter 2 letters ahead or 8 or 13 letters ahead should not make any difference; it is the same task. This failure reveals a basic lack of an ability to apply a deductive rule. Their memory tests show a similar frequency effect—99% accuracy in recalling the birthday of famous and frequently mentioned figures (e.g., Jeff Bezos) versus ~23% for infrequent ones (McCoy et al. 2023, pp. 31–32). Thus, LLM output quality tracks mention frequency, reflecting the frequency bias inherent in the architecture. For strategists, this means that insights about well-known firms or patterns are often reliable, whereas reasoning about obscure competitors or novel conditions invites errors and hallucinations.

As McCoy et al. (2023) show, the core failures in tasks like rot-13 persist even when tokenization effects are carefully controlled (including cases where input and output strings are matched in token length and character structure), so the problems are not merely the result of how words are converted into tokens; even if they were, however, it remains a problem for LLMs. More importantly, performance results systematically track the probability in the training distribution rather than its logical or computational complexity. Closely related transformations with identical tokenization properties succeed or fail dependent upon corpus frequency, indicating that the errors arise from architectural structure rather than from surface-level tokenization issues alone.

Strategy frequently requires inference from sparse, unusual, or novel data—precisely where LLMs degrade. Even small syntactic changes ("12 × 3" versus "multiply 12 and 3") reduce reliability. Deductive failures reveal that LLMs improve, typically only when prompts match previously seen phrasings (Lampinen et al. 2024, Seals and Shalin 2024).[5] Benchmarks, such as ReClor, LogiQA, AR-LSAT, ProofWriter, ADIE, and BIG-Bench Hard, confirm that even 2025 frontier models (GPT-4.2, Gemini 2.0 Ultra, Claude Opus 2025, and DeepSeek-R1) underperform humans on multistep, rule-governed deduction. LLMs excel at interpolative deduction—cases well represented in training—but fail on generalization requiring stable rule schemas or symbol manipulation. They articulate the center of mass of observed deductive patterns rather than executing deduction as a rule-based procedure, typically scoring ~30% below human performance across formal benchmarks. Most deductive reasoning benchmarks show that LLMs fall about 20% behind human performance. See Online Appendix C for details. LLM success here would be, perhaps, a human-equivalent score on more than half of the benchmarks (after eliminating the poorly constructed tests).

5.4. Causal Reasoning in Systems

Causal reasoning in systems refers to the ability to think through complex cause-and-effect structures and feedback loops within a business environment (Senge 1990, Sterman 2000, Repenning 2002). Strategists rely on causal mental models to anticipate second-order and third-order effects—such as how a price cut can increase demand, provoke competitive retaliation, alter channel incentives, or strain operations (Sterman 1989, Gary and Wood 2011). A long line of research shows that effective strategic judgment depends upon constructing accurate causal representations of the environment, whereas inaccurate causal beliefs generate persistent biases and policy failures (Forrester 1961, Kaplan 2008, Gavetti 2012). By contrast, current LLMs do not possess genuine causal reasoning but instead, approximate causal discourse (Pearl 2000) through pattern imitation (Marcus 2022). Scholars in strategy and operations have repeatedly emphasized that machine learning systems fail to infer causal structure and therefore, cannot support the kind of counterfactual reasoning required for robust strategy formation (Rahmandad and Sterman 2012, Felin and Holweg 2024). As several authors argue, LLMs' outputs can mimic causal explanation but "lack the underlying model of the world that gives causal statements their force" (Marcus and Davis 2020, p. 112; see also Zečević et al. 2023; Wu et al. 2024; and Zhou et al. 2024a, b), making causal reasoning an area of persistent concern and frequent critique in evaluations of AI tools for strategy. The core obstacle remains simple; understanding the world through the lens of association patterns in text only gets you so far.

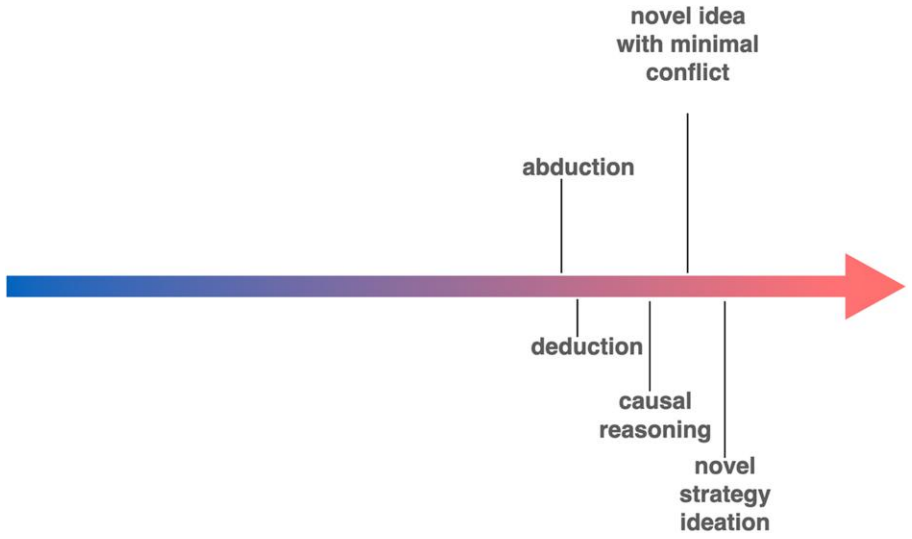
Empirical benchmarks from 2020 to 2025 show that LLMs perform well at associational causal reasoning (e.g., identifying likely causes from static text) but degrade sharply when required to evaluate interventions or counterfactuals. Benchmarks, such as WIQA, CausalQA, COPA, TÜlu Causal, CounterfactualNLI, CausalBench, and the BIG-Bench Causal Reasoning tasks, reveal a consistent pattern; models excel when the causal relation is common or explicitly stated, but they fail when causal structure must be inferred, when background knowledge is sparse, or when reasoning involves complex multivariable interactions.

Current flagship models still fall far short of human-level counterfactual and interventional reasoning, a gap indicative of a familiar structural issue—the mean articulation problem.

5.5. Novel Idea Generation with Minimal Conflict

We are now deep into "the red zone" (see Figure 4). We know that very rare ideas are relegated to the long tail of near-zero probability (Bridgelall 2024; Zhang et al. 2024). Truly novel ideas are then mostly hopeless for LLMs, and any form of conflicting information reduces their chances for success even further. This is the "transformational creativity" from Boden (2004) which requires "changing the rules of the conceptual space, so

Figure 4. (Color online) Difficult Tasks for LLMs



that ideas can be generated which were impossible before" (Boden 2004, p. 6). This requires, in other words, that existing accepted knowledge be contradicted—albeit only limited knowledge at this task level. When an LLM attempts to generate a genuinely novel idea that conflicts with some portion of a large body of consensus training data, lacking any competing training data about the novel idea means that there is no weight of evidence to "pull" the system away from the consensus information. The "mean articulation" architecture explains why LLMs fail to generate new ideas that contradict a well-established consensus fact. The same structural brittleness that prevents robust causal reasoning also makes paradigm-breaking novelty unattainable (Figure 5).

5.6. Novel Strategy Ideation

Novel strategy ideation concerns the generation of business models or opportunity sets that depart from prevailing industry logics and dominant designs (Schumpeter 1934, Burgelman 1983, Kim and Mauborgne 2005). As such, this is not a simple recombination of existing ideas. Sustained advantage typically requires a departure from standard ideation because VRIO resources or configurations are almost never identifiable through local search alone (Barney 1991). Felin and Zenger (2017) emphasize that strategists act as theorists, formulating causal hypotheses about value creation that may contradict existing evidence or industry beliefs. This is a stronger instance of the Boden (2004) "transformational creativity," which requires contradicting widely established training data ("changing the rules of the conceptual space"). This task is more difficult than constructing a novel idea that only involves minimal conflict or standard strategy ideation because truly novel high-rent yielding strategies are almost always hidden behind a wall of countering consensus (McBride et al. 2024). This is why this task sits farther right on the task continuum.

The architecture of LLMs is mismatched to this task. They are biased toward high-frequency text continuations and away from rare or counterintuitive propositions (Bender et al. 2021, Zhang et al. 2024). Empirical work shows that when truth conflicts with

Figure 5. (Color online) The Most Difficult Tasks for LLMs

high-frequency beliefs, models tend to reproduce the latter (Lin et al. 2022). Creativity studies find that LLM-generated ideas are appropriate recombinations of common elements but that they converge toward a narrow set of themes and avoid genuinely unusual structures (Binz and Schulz 2023, Wenger and Kenett 2025). This makes them strong interpolators but poor extrapolators to outlier theories—the mean articulation problem.

Research in entrepreneurship reinforces this view. Machine learning algorithms can match or outperform angel investors on average judgments, yet they systematically miss extreme winners that depend on recognizing outlier opportunities (Blohm et al. 2022). Studies on AI-supported venturing similarly find that generative tools assist refinement and communication but rarely originate "rogue" entrepreneurial ideas that appear implausible under current evaluative frames (Chalmers et al. 2021). Organizational creativity research shows that AI tools increase the volume of ideas but primarily by amplifying existing patterns rather than reframing problem spaces (Amabile 2020, Jia et al. 2024, Farrell et al. 2025). Computer science work on scientific hypothesis generation reaches similar conclusions. Systems, such as MOOSE, generate plausible hypotheses only within well-represented regions of prior scientific discourse, effectively interpolating within known spaces (Yang 2024, Yang et al. 2024).

Accordingly, LLMs are useful for simple strategy ideation (farther left on the continuum)—listing business model variants, surfacing analogies, or recombining familiar strategic elements (Hamel and Prahalad 1994, Gavetti and Rivkin 2005, Csaszar et al. 2024). They extend exploitation and local search (March 1991, Gavetti 2005). But, because they articulate documented ideas, they are unlikely to generate VRIO strategies, radical business model innovations, or counterconsensus value theories (Barney 1991, Felin and Zenger 2017), even when prompted to do so. Novel strategy ideation thus remains among the tasks least suited to LLMs; it relies on abductive theorizing, causal reframing, and deliberate violation of prevailing data patterns—capacities that, at present, remain more or less distinctively human.

At present, there are no established benchmarks that directly evaluate the ability to generate truly novel ideas or to perform novel strategy ideation in the sense required by strategic management theory. Existing creativity and ideation benchmarks—whether in psychology (e.g., TTCT-like tasks), computer science (e.g., CreativityPrism or PACE), entrepreneurship (e.g., new-product or opportunity-recognition studies), or scientific hypothesis generation (e.g., MOOSE or Scideator)—all assess forms of combinatorial creativity where the system recombines familiar conceptual elements into outputs that remain within the boundaries of existing knowledge and evaluative frameworks. None are designed to reward or even detect the kind of truly novel creativity that involves breaking, revising, or transcending prevailing conceptual or industry logics while still producing coherent value-creating ideas. As a result, we currently lack empirical tools capable of measuring whether any system—human or machine—can reliably generate deeply novel yet minimally conflicting ideas or develop genuinely new strategic theories that depart from dominant designs.

5.7. Decision Making Under Uncertainty, Long-Range Planning, and Generating a Novel Idea When There Is Maximum Conflict

The final three tasks will now be explored.

5.8. Decision Making Under Uncertainty

Often, decisions must be made without complete information. Strategic decisions typically fall on the Knight (1921) "uncertainty" side of risk versus uncertainty dichotomy, where situations have unknowable probabilities because for example, they involve ambiguous signals, incomplete data, contested interpretations, or evolving contexts. Behavioral strategy scholars have emphasized that such decisions require the integration of analytical reasoning, intuition, and judgment (Simon 1957, Powell et al. 2011). Entrepreneurs similarly exercise what Kirzner (1973) called "alertness," which is the ability to notice overlooked possibilities in environments where information is sparse or equivocal. Additionally, as Foss and Klein (2012) argue, judgment under uncertainty is ultimately an act of conjecture, not calculation.

The central importance of this cognitive task for strategy derives from the fact that organizations rarely face environments characterized by stable distributions or known outcome likelihoods. High-stakes choices, such as entering an emerging market, adopting a novel technology, or responding to a competitor's unexpected move, force decision makers to project forward from fragmented evidence, to reason about causal mechanisms that may not yet be observable, and to commit resources despite ambiguity (Gary and Wood 2011, Gavetti 2012). Judgment under uncertainty, therefore, often requires abductive leaps, counterfactual reasoning, and the ability to form internally coherent beliefs even when the available data cannot determine a single best answer. A common example is an executive deciding whether to pursue a nascent technological platform. Despite limited evidence and contradictory expert opinion, the strategist must evaluate alternative causal stories (e.g., whether early adoption will confer network advantages or whether the technology will fail to mature) and commit the firm to a path that may be irreversible.

Within the mean articulation architecture discussed above and in the work of McCoy et al. (2023), current

LLMs can provide extensive assistance with decision making under uncertainty to the extent that some of the problem can be decomposed into tasks that fall on the "easy" left side of the task continuum—summarizing known patterns, aggregating documented evidence, or articulating high-frequency interpretations. But, they falter precisely where uncertainty becomes genuine in the Knightian sense—when the relevant information is absent from the training corpus, when causal mechanisms are ambiguous or undocumented, or when the situation requires abductive leaps that contradict prevailing consensus (Lin et al. 2022, Bridgelall 2024). Because LLMs are structurally biased toward high-frequency continuations, they amplify established beliefs rather than entertain contrarian hypotheses, making them particularly ill suited for strategic uncertainty, where rare insights, nonobvious causal models, or low-frequency signals often matter most (Gavetti 2012, Felin and Zenger 2017). An LLM evaluating whether to invest in an unfamiliar frontier market, for example, will gravitate toward documented cases of similar markets, reproducing consensus narratives rather than generating the speculative causal conjectures that a strategist must consider. Thus, LLMs may serve as powerful aides for organizing known information, but they cannot yet substitute for human cognitive capacities, like more difficult cases of abduction, as well as causal theorizing or counterconsensus judgment required for decision making under true uncertainty.

Currently, we lack any benchmark that measures the ability to generate judgment under radical, domain-changing uncertainty—the very form of uncertainty relevant to strategy and entrepreneurship.

5.9. Long-Range Planning

In strategic management, long-range planning refers to the deliberate process by which firms articulate desired future states and develop coordinated actions to reach them over multiyear horizons. Classical accounts emphasize the need to envision plausible future environments; identify long-term goals; and integrate resource allocation, capability development, and environmental positioning across extended time frames (Ansoff 1965, Schoemaker 1995). The activity is inherently judgment laden because firms face structural uncertainty about markets, technologies, competitors, and regulatory trends; strategic plans must, therefore, accommodate ambiguity, shifting constraints, and emergent opportunities (Mintzberg 1994, Makridakis et al. 2010). Long-range planning also requires constructing and maintaining a coherent causal model of how competitive advantage will be created and preserved over time—an inherently theory-driven process in which strategists must project future scenarios, anticipate responses, and design adaptable paths (Teece et al. 1997, Gavetti 2005).

In computer science, however, long-range planning refers to the ability of an algorithm or agent to execute extended, multistep sequences of actions in pursuit of a defined objective. This is a skill that has also benefited from CoT reasoning but is still often seen as one of the most difficult challenges and one often discussed by software engineers (e.g., LeCun 2022a, b; Bengio 2024; Kambhampati et al. 2024; Valmeekam et al. 2024). Studies by Valmeekam et al. (2023) and Laban et al. (2025) show that unaided models often drift, contradict earlier steps, or lose coherence the longer the chain of actions is. LLMs can generate step-by-step outlines if prompted ("plan a marketing campaign" or "design a travel itinerary"), but they lack stable mechanisms for extended goal execution, maintenance, and updating. Although more memory helps, models often ignore or misprioritize earlier steps (Hsieh et al. 2024, Bai et al. 2025). Researchers emphasize that autonomous planning remains unreliable unless models are scaffolded with external tools, symbolic planners, or iterative reasoning frameworks (Kambhampati et al. 2024). Even recent reasoning-specialized models (e.g., DeepSeek-R1 and OpenAI o1) show improvements in short-horizon reasoning but continue to falter on open-ended, long-duration tasks. Thus, long-range planning sits at the hardest end of the continuum.

Simple, strategic, long-range planning in very stable, known, predictable environments is mostly straightforward for both humans and machines, but it is a rarity. Realistic long-term planning typically confronts a wide assortment of environmental uncertainties and requires precarious causal reasoning into the future, which limits its success. Agentic or algorithmic long-range planning in LLMs is famously difficult because minor flaws in the the construction of a series of tasks and the execution compound to collapse the effort.

Although computer science includes numerous benchmarks nominally categorized as "long-range" or "long-horizon" planning—ranging from PDDL classical-planning domains (Blocksworld, Logistics, and Rovers) to long-horizon robotics suites (Meta-World, CALVIN, and ManiSkill2), exploration-heavy RL environments (Montezuma's Revenge and Pitfall), tool-use settings (WebArena and BrowserGym), and language-based planning evaluations (PlanBench and ALFWorld)—the empirical pattern is consistent; LLMs perform poorly on all tasks that require genuine long-horizon planning (Asai et al. 2022; Hao et al. 2023; Valmeekam et al. 2023, 2024; Kambhampati 2024). Older (symbolic) classical symbolic planners routinely achieve 90%–100% success in PDDL domains, whereas LLM-based planners typically reach only 20%–60% success, with performance collapsing as horizon length grows. In robotics benchmarks, state-of-the-art RL agents reach 60%–90% success on long-horizon tasks, whereas LLM controllers remain in the 10%–40% range. This is one area where progress has

gone backward compared with "good old fashioned" symbolic AI. In sparse-reward RL environments, like Montezuma's Revenge, humans and top RL systems achieve near-perfect scores, but LLM agents remain below 10%. Even in language-based planning tasks, such as PlanBench or ALFWorld, the best 2025 LLMs plateau around 50%–65% on multistep plans, with errors compounding rapidly after 5–10 steps. Across all domains, no benchmark shows LLMs achieving robust, reliable long-range planning. Current systems are effective at short, local action sequencing but not at maintaining or executing coherent plans over extended horizons, especially under uncertainty or multistep dependencies.

5.10. Novel Ideas with Maximum Conflict

We arrive at the end of the continuum and what is the most unapproachable task for an LLM—the generation of a truly novel idea (i.e., an idea that does not yet exist in the world, much less the training data) that conflicts with multiple consensus facts. This challenge contains all of the problems of generating a novel idea discussed earlier but multiplies the difficulty by multiplying the number of contradicted ideas, eliminating any chance of succeeding at this task. This can be thought of as "extreme transformational creativity," an extreme case of the Boden (2004) category. Of course, extreme transformational creativity is no easy task for humans either, but it is at least occasionally done. This is the stuff of innovative product breakthroughs, true scientific advances, and the rare Kuhnian paradigm shifts. It is silly and easy to state a maximally contradictory novel idea like "planets in our solar system are comprised entirely of Swiss cheese," but it is incredibly difficult to come up with one that is actually true and provides greater explanatory power. These are arguably the most interesting and important of achievements in science, strategy, entrepreneurship, and any social advances in general. It sits on the far end of a continuum of task difficulty for LLMs because it is the perfect antithesis of a mean articulation system whose responses respect the frequency of occurrence in the training data. Finding truth outside of any articulated ideas, a truth that fits coherently with important existing knowledge but rejects many well-established pieces of it, is a peculiar event that inspires researchers in the fields of philosophy of science, innovation, "blue ocean" strategies, and entrepreneurship.

There are no existing benchmarks—in computer science, psychology, economics, strategy, or creativity research—that measure the ability to generate truly novel, although still feasible, ideas that simultaneously contradict multiple entrenched, consensus facts yet remain true, coherent, and generative.

We have laid significant groundwork, first in an examination of LLM architecture basics and then in identifying an increasingly difficult range of tasks. There are many other tasks that could be added to this continuum of course and various possible reorderings, but the basic trajectory is correct for the simple reason that as we move farther down the scale, we move farther away from essential "mean-articulating" architecture of LLMs. The same design that made possible the incredible advances of LLMs also hinders their aptitude on certain tasks. See the complete task range in Figure 6 and the fuzzy boundary, which is denoted by the dashed line in Figure 6, that separates the area of LLM competence from the area of LLM incompetence. The majority of activities fall inside the zone of LLM competence, which should be should be good news to practitioners.

Next, we will look in more detail at this most difficult challenge for LLMs—the creation of a novel idea that conflicts extensively with existing knowledge—and explain why this is disproportionately important for strategy before discussing additional implications for strategic decision making.

6. The Hardest Task for Strategy

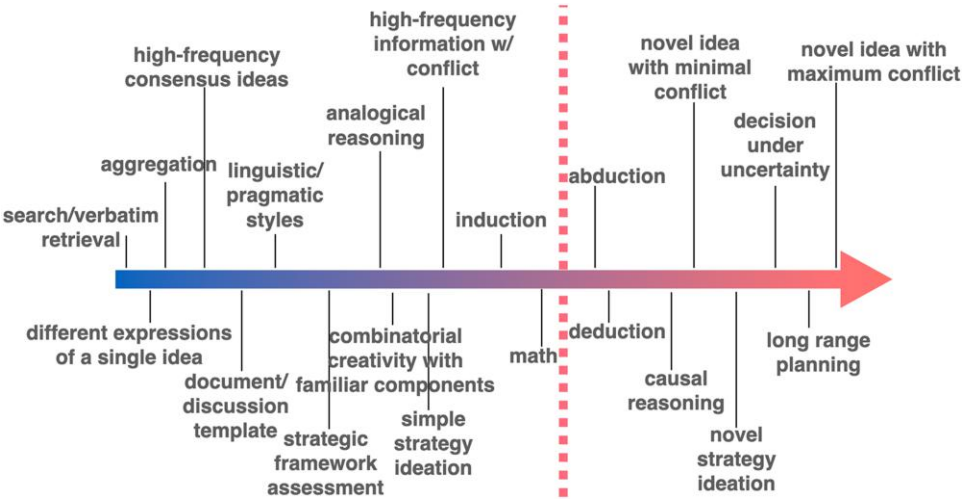
6.1. The iPhone Composition Decisions

Apple's 2007 introduction of the iPhone provides a vivid example of a strategic decision that defied industry consensus and illustrates why a "mean articulation machine" would have been incapable of envisioning the same breakthrough. In the mid-2000s, dominant assumptions about mobile phones were firmly established. First, mobile design conventions held that input required physical keys or a stylus. Virtually all smartphones featured hardware keyboards (e.g., BlackBerry's thumb keypads) or resistive touchscreens operated with a stylus. These fixed plastic buttons were seen as essential for typing and navigation. A comment from early 2007 captured the prevailing sentiment: "[N]o [QWERTY] keyboard? … [using] some weird finger tapping on a screen [is] crap." The consensus was that purely touch-based input would be error prone and unwelcome to users. Even Microsoft's CEO Steve Ballmer laughed that "no one would want to type on glass" when he first saw the iPhone's design (Isaacson 2011).

Second, batteries should be removable. Circa 2006, every popular phone had a swappable battery. Consumers and carriers alike expected the convenience of carrying spares and replacing batteries as needed. A sealed, nonremovable battery was almost unthinkable—a "jarring design choice" at the time. Apple's own engineers knew that this would flout consumer expectations, yet they intentionally abandoned the removable battery in the iPhone to enable a thinner, sturdier device (Isaacson 2011).

Third, a mobile operating system should be appropriately denuded for a small device. It should be

Figure 6. (Color online) The Full Continuum of Tasks for LLMs and the Approximate Boundary of Competence



lightweight and specialized (e.g., Symbian, PalmOS, or Windows Mobile). The notion of running a desktop-class operating system on a phone seemed far fetched because of CPU, memory, and power constraints. Industry experts assumed that a phone had to sacrifice computing power for battery life and size. Apple's decision to equip the iPhone with only a slightly scaled-down version of OSX—essentially a desktop-class OS in a phone—ran against these constraints. It meant treating a phone more like a handheld computer, an approach that no rival had attempted. This violated the dominant belief that mobile devices could not handle PC-like software; Apple's gamble was that a richer OS would enable transformative apps and web experiences, even if it taxed the hardware (Isaacson 2011).

Fourth, screens should be plastic, not glass. Leading up to 2007, phone screens were typically plastic for shatter resistance, even though they scratched easily. A glass screen on a phone was considered too fragile. Yet, Apple insisted on a glass multitouch display, prioritizing optical clarity and scratch resistance at the risk of fragility. After Jobs found that his iPhone prototype's plastic screen had scratched in his pocket, he famously demanded a hardened glass solution—even though such glass "didn't exist yet" and mass producing it in time for launch seemed impossible (Vogelstein 2013). Apple pushed Corning to deliver a new Gorilla Glass within six months, a Hail Mary engineering gamble that no conventional analysis would have recommended. This choice directly defied the norm of using safer plastics.

Fifth, mobile networks needed to be third generation (3G). By 2007, 3G wireless was emerging as the standard for high-end smartphones; competitors were beginning to offer 3G data for faster email and web browsing. It was consensus that any premium device should utilize the fastest network available. The original iPhone, however, launched with only the slower EDGE (2.5G) support, which many observers criticized as a mistake. Apple had to explain that 3G chipsets were too power hungry and space consuming at the time (Thompson 2012). In other words, Apple knowingly contradicted the industry's 3G-first mantra, betting that consumers would accept slower cellular speeds in exchange for the iPhone's other advantages (and that it could use Wi-Fi for high data needs). This was a low-probability choice given the era's expectations; reviewers even called EDGE "the iPhone's Achilles' heel" at launch, underscoring how against the grain this decision was (Mossberg 2007).

Each of these design decisions individually violated a prevalent consensus belief. Apple did not just take one contrarian bet; it took all of them at once. A single risky move (say, removing the keyboard) might be explained away as an eccentricity, but the iPhone represented a constellation of contrarian choices. It was, in effect, multiple low-probability ideas combined into one product vision. Prior to the iPhone's debut, no expert or analyst had publicly speculated that a phone would omit a keyboard and stylus, use only a touchscreen with multitouch gestures, seal the battery, run a desktop OS, use an all-glass front, and eschew 3G—all at the same time. The concept was beyond the edge of the industry's collective imagination.

Notably, even within Apple there was hesitation. Tony Fadell (2012) recalls that early iPhone prototypes included one version with a hardware keyboard as a hedge against the risk that a software keyboard might fail. Only after internal debate did Jobs and his team commit fully to the radical design. This underscores how unusual the pure-touch approach appeared; it surprised seasoned Apple engineers, not just outsiders.

Market reactions in 2007 mirrored the consensus biases of the time. Many observers simply could not

accept that Apple's no-keyboard, finger-only interface would be usable. Tech commenters scoffed that a touchscreen phone was "massively disappointing," predicting "obvious and pretty major problems" with typing and usability. Pundits questioned the battery life ("5 hour battery life … ARE YOU KIDDING, this is a phone not a laptop" one said) and the lack of an established mobile OS or app ecosystem. In short, the initial consensus was that Apple was making a huge mistake by defying so many conventions. Reinforcing incumbents' views, Nokia's internal analysis of the iPhone noted its "lack of physical keyboard and high price" as weaknesses that would limit its appeal, even as they acknowledged the device's impressive UI innovations (Surowiecki 2008). BlackBerry's leadership similarly dismissed the iPhone, convinced that business users would never give up physical keys. This collective incredulity from experts was the majority opinion, and Apple had deliberately chosen to contradict it.

Imagine prompting a state-of-the-art language model in 2006 (using all relevant data available at that time) to "design an innovative new more powerful mobile phone." The model's suggestions would inevitably gravitate toward the center of mass of 2006-era thinking. It might have produced a design similar to a BlackBerry or Nokia communicator—perhaps a somewhat improved candy-bar phone with a physical QWERTY keyboard, maybe a stylus for precision input, a removable battery, and support for 3G networks. It would not have drawn the improbable features together to envision anything like the iPhone because each of those features was an outlier in the data.

As noted, LLMs not only fail to suggest counterconsensus ideas, but also, they are biased against them by construction. This leads to a "herding" effect, where everyone gets the same advice drawn from the same best practices (Felin and Zenger 2017, Farrell et al. 2025) if they are not "counterprompted." If all firms in 2006 believed that a successful phone must have a keyboard and removable battery, an AI simply regurgitating that conventional wisdom would reinforce the herd mentality. LLMs are ill suited for "rare events" or bold one-off decisions that depart from established patterns. They cannot say: "All the data says X, but despite that, Y might be true." At best, they can reflect minority opinions present in the training data, but in the case of the iPhone's features, there essentially were none. The necessary ideas (multitouch UI, soft keyboard, etc.) were so novel that they appeared only as nascent research concepts or not at all in mainstream discourse. Consequently, an LLM would have had no grounds to combine these into a single vision.

Strategic management research emphasizes the value of contrarian insight. The Barney (1991) VRIO framework holds that a strategy yields sustained advantage only if it employs resources or decisions that are valuable, rare, inimitable, and nonsubstitutable. By definition, "rare" moves will not be drawn from the pool of common ideas; they require creative foresight against prevailing logic. The iPhone exemplified a VRIO strategy; its design was valuable to consumers, rare in the industry, hard for rivals to imitate quickly (because of Apple's unique integration of software and hardware), and supported by the organization. It took a different mindset to assemble these rare choices into a coherent product. As Gavetti (2012) noted, breakthrough strategic decisions often come from leaders' courage to break from the herd and envision possibilities that contradict experiential learning. LLMs lack the epistemic vigilance and creative doubt required to say that "the conventional wisdom might be wrong here" (Gavetti 2012). In the iPhone case, a human had to apply nonconsensus reasoning—seeing, for example, that a multitouch UI could work if done right, even though all prior evidence suggested that touch keyboards were awful. That leap of faith is alien to an LLM's articulations that prioritize high-frequency-exposed data, as we have seen.

In sum, the original iPhone case underscores the limitations of an LLM as a strategic ideation partner for this kind of end-range task, and this case study thus serves as a caution; if a firm relies on LLMs for strategic guidance, it will miss this category of paradigm-shifting ideas that create real competitive advantage. Well-documented responses can reproduce and even refine past knowledge, but they cannot easily invent the future. The iPhone's genesis highlights the indispensable role of human imagination and contrarian thinking in innovation—capabilities that, at least for now, lie beyond the reach of large language models' consensus-bound architectures.

6.2. Novel Theories

LLMs' inability to compose genuinely novel ideas applies to generating genuinely novel theories as well. The theory-based view (Felin and Zenger 2017) emphasizes that economic actors, like scientists or strategists, are not simply passive observers but active theorists who construct causal models about how they might create and capture value. Success here arises not from averaging existing observations but from proposing explanations that go beyond the available data and established consensus. Gavetti (2012) similarly stresses that strategic progress depends on theorizing under uncertainty, where leaders exercise imagination and foresight to envision possibilities that cannot be directly inferred from experience. This theoretical leap is what allows actors to identify rare, valuable insights that others overlook. Such abductive daring rather than, for example, iterating a product idea from customer feedback is essential for innovative advances

(Felin et al. 2020). The same is true of innovative scientific advances, like Einstein's special relativity, where progress depended on discarding entrenched assumptions and articulating a radically new theory of space and time (Einstein 1952).

Einstein's 1905 development of special relativity provides another clear case of a task at the far end of the continuum (Einstein 1952). The prevailing scientific paradigm at the turn of the twentieth century was deeply rooted in Newtonian mechanics, which assumed absolute time, absolute space, and a universal ether through which light propagated (Whittaker 1951, Holton 1973). Physicists, such as Lorentz and Poincaré, had introduced mathematical corrections to accommodate anomalies, like the Michelson–Morley experiment, but their formulations preserved the ether framework and the absolute nature of temporal and spatial reference (Pais 1982). The consensus scientific corpus overwhelmingly reinforced Newton's framework, with only marginal dissent. An LLM trained on this literature would have treated Newtonian assumptions as overwhelmingly likely continuations, much as today's LLM models amplify cultural myths when they appear frequently in training data (Lin et al. 2022).

What made Einstein's insight radical was not a modest recombination of existing concepts but the rejection of multiple entrenched assumptions simultaneously (Einstein 1952). Special relativity denied the absoluteness of time, rejected the ether, and treated the speed of light as constant across inertial frames—all propositions that contradicted dominant consensus (Einstein 1952). The weirdness of the theory is that unlike all other objects whose speed varies with your speed (i.e., a car traveling at 50 mph only appears to be traveling at 20 mph if you are traveling next to it at 30 mph), the speed of light is absolute no matter how fast you are traveling next to it. A language model trained on 1905-era discourse could not have produced Einstein's radical reconfiguration.

The direct implication for creative breakthroughs in strategy or science should now be clear; LLMs, in their current architecture, will never make scientific or strategic breakthroughs because they are designed to articulate the most widely established ideas. Their architecture, in other words, is antithetical to contradicting widely established knowledge. One might reply: "But, you can prompt an LLM to specifically reject any widely accepted theory you choose by simply telling it to do so." This is true. But, then how will you know which theories must be rejected in what ways to allow for the as-yet-unknown breakthrough theory that replaces it? One cannot.

Strategic theorists and historians of science alike emphasize that such breakthroughs require the capacity to hypothesize that reality might work differently from both empirical appearances and what prior theories suggest. Einstein's insight depended not only on mathematical reasoning (Einstein 1952) but also, on imaginative thought experiments, such as chasing a beam of light, which revealed contradictions invisible within the consensus paradigm (Holton 1973). LLMs, by contrast, cannot perform such counterconsensus hypothesis generation; they have no embodied model of physical processes, and their probabilistic architecture structurally suppresses precisely the kinds of low-frequency combinations that define revolutionary science. Thus, just as they could not have "invented" the iPhone, they could not have "discovered" special relativity; both cases illustrate that breakthroughs emerge from rare, contrarian reasoning, not from statistical averages of what has been said before.

7. The Limit of LLM Utility for Strategic Decision Making

The knotted question about whether LLMs can be useful for strategic decision making has been approached by decomposing it into a range of strategic cognition tasks. The general answer to the puzzle is now straightforward. To the degree that an optimal strategic decision lies on the right end of the continuum, the less appropriate it would be to deploy an LLM. On the right side are tasks that demand counterconsensus reasoning, robust causal inference, abductive hypothesis generation, deductive consistency, and the ability to sustain long-range plans. Here, the evidence is unambiguous; models falter because their essential design architecture is optimized for association-based linguistic reproduction biased toward high-frequency text, not epistemic vigilance or insight. They misinterpret or smooth over anomalies (Lin et al. 2022), default to high-frequency continuations (Bridgelall 2024, Zhang et al. 2024), and often produce the illusion of reasoning rather than genuine problem-solving (Apple Machine Learning Research Team 2025).

The more specific result is that generating a novel idea that violates multiple established (documented) consensus beliefs is beyond the current reach of LLMs. Strategic or theoretical breakthroughs, like Einstein's relativity or Apple's iPhone, require rare leaps that violate multiple entrenched beliefs—tasks squarely in the "red zone" of the continuum. This would then preclude the construction of a Barney (1991) VRIO-type strategy because the "rare" requirement (at least this specific type of rare idea) is unapproachable. Things are not much better in generating a novel idea that does not exist in the training data that contradicts only one well-established belief. But, at least there are fewer obstacles.

This issue has significant implications for strategic decision making because successful strategies often emerge from contrarian insights that recognize opportunities that the majority underestimates, rejects, or

misinterprets. For example, disruptive innovations, countercyclical investments, or bold market entries usually involve going against prevailing assumptions. If LLMs systematically "average" toward consensus, they will tend to recommend safe, conventional strategies rather than bold or contrarian ones (Raisch and Krakowski 2021). This mirrors their failure in the TruthfulQA benchmark; they echo what is most often said rather than what is true. For managers and entrepreneurs, this means that LLMs may be excellent tools for synthesizing mainstream perspectives or mapping the current strategic landscape, but they are poorly suited for identifying or justifying outlier strategies. In fact, their consensus bias could reinforce herding behavior—where organizations converge on similar "best practices"—and discourage the exploration of unique, high-risk, high-reward moves. Strategic breakthroughs, like Apple's iPhone, require precisely the kind of reasoning that violates consensus. As a result, the role of the human strategist remains indispensable; top-level management must integrate LLM-generated summaries of conventional wisdom with their own capacity for theoretical reasoning, contrarian thinking, and courage to break from the herd (Gavetti 2012, Shepherd and Majchrzak 2022).

The distinction between familiar tasks easy for LLMs and exploratory tasks that requires extrapolating beyond the training data echos the distinction of March (1991) between exploitation and exploration. That distinction highlights the organizational trade-off between refining existing routines versus venturing into uncertain, riskier domains. Adner and Levinthal (2008) extended this framework by showing that exploration involves not only "doing" through experimentation but also, "seeing" differently—reframing perceptions and developing new mental models of what might be possible. Per March (1991), the benefits of exploration include innovation and novelty, adaptability in the form of building capacity to adapt to changing environments, learning under uncertainty, the long-term survival that comes from discovering new opportunities, and avoiding competency traps. The implication here for strategy is that machines exploit through interpolation rather than explore and extrapolate.

But, of course, any given strategic decision might (or might not) rely on an assortment of more appropriate uses of LLMs on the difficulty continuum. The discussion so far has focused more on the limitations of LLMs. But, it is worth emphasizing the enormous assistance of LLMs on the majority, easy end of the task spectrum where they can be used as the lowest-paid research assistant that one ever had that never requires sleep. The left side of the continuum highlights the immense value that LLMs already offer. Search over the world's compendium of knowledge, aggregation, analogical reasoning, strategic framework assessment, paraphrasing across syntactic forms, pragmatic conversational-style replication, combinatorial creativity, and high-quality consensus summarization is well within their domain. For strategists, this means that LLMs excel as tireless research assistants: synthesizing large literatures, formatting proposals, drafting reports, imitating professional styles, and mapping the terrain of well-established knowledge. Even in breakthrough contexts, such as the iPhone's glass screen gamble, an LLM could not have generated the idea, but it could have provided invaluable support once the hypothesis was on the table—for example, by rapidly summarizing the state of glass technology, analyzing manufacturing trade-offs, or surveying comparable engineering challenges.

Thus, the continuum framework clarifies the role of LLMs; they are indispensable at the left end for speed, fluency, and breadth but less reliable at the right end, where novelty, abduction, deduction, and causal reasoning dominate. For strategy, the implication is straightforward; the further a decision lies toward the right end of the continuum, the more it requires human theorizing, judgment, and contrarian courage. The further it lies toward the left, the more it can be safely handled by LLMs. In knowledge industries that depend upon efficient transmission of consensus knowledge, this capacity is disruptive and transformative.

This continuum offers an orthogonal and inclusive perspective on LLM capabilities, overlapping with and complementing the standard AI benchmarks in computer science with a broader view of what these models can and cannot do. Other researchers will no doubt build more strategy-specific benchmarks and expand and edit this continuum of skills in future work as part of the process of charting LLM progress.

The central goal of this paper has been not only to examine the applicability of AI (specifically LLMs) to strategic decision making but in doing so, to also grapple with the first of the fundamental questions about LLMs themselves: "Where exactly is the boundary of what can currently be extrapolated by LLMs?" The discussion here has been an extended answer to this question. The broad answer is that the same feature that currently provides such impressive results limits the extent to which they can extrapolate beyond the training data.

But, we can now shed some light on the other questions as well: "Can the boundary of LLM extrapolation extend to genuine knowledge breakthroughs?" Given what we know about end-range novel ideas and the challenges of contradicting existing data in paradigm-shifting ideas, the current architecture makes this impossible. Scientific breakthroughs are stated in text, but that does not mean that they can be discovered through textual analysis. Alternative architectures custom built for knowledge extension in more focused

domains, like biological protein combinations tackled by AlphaFold2, will proliferate in the future.

There were two further fundamental questions. "How far can the extrapolation boundary be expanded?" "Can they extrapolate beyond their training data into undocumented knowledge?" By laying out a continuum of LLM skills, we see at least the sketch of an answer to these questions. In many complex domains like strategy decision making, where optimal answers are often not clearly written down in any text, the question of how far can LLM extrapolate beyond their training data is relevant. Optimists argue that with enough data and larger models, LLMs might eventually push beyond the boundaries of documented and undocumented human knowledge. They envision that the extrapolation boundary of LLMs could expand to encompass not only everything that humans currently know but even to make inferences that exceed what any human has yet discovered (Kurzweil 2022, Bubeck et al. 2023, OpenAI 2023, Uehara 2025). This optimistic perspective, buoyed by recent rapid advances, suggests that increasingly powerful models might exhibit emergent capabilities tantamount to creativity or genuine insight, effectively using the "ripples"—the hidden patterns in known text—to infer the presence of uncharted ideas in the larger sea of possible knowledge. Simple patterns in text associations are all that one needs to make knowledge breakthroughs.

The conclusions above align more with a conservative view. The current extent of extrapolation is limited by the mean articulation architecture. Additionally, when it comes to strategic decision making and other complex tasks that require insight beyond what is written, LLMs are currently not close to eclipsing human experts. A significant portion of strategic knowledge in business (and human knowledge in general) remains undocumented, existing only in the minds of practitioners or in tacit understandings built through lived experience. Strategic decisions often hinge on intuitions, hunches, insider information, trust-based relationships, and contextual nuances that have never been codified in any database or text (and perhaps cannot be fully captured in words).

It would be pure speculation to estimate what percentage of human knowledge and know-how is undocumented or undocumentable. But, much implicit knowledge in the form of neuromuscular skill, interpersonal social intuitions, and interaction expectations (what Searle 1983 referred to as "the background") is not only not documented but, often impossible to describe (see Polanyi 1966). Similarly, no amount of internet-scale text training can grant an AI access to a firm's confidential plans, a CEO's gut feeling about a market shift, or the unspoken social dynamics within an industry. By definition, all of this is absent from text-based training corpora. Given these realities and the fact that a significant portion of human knowledge and skill is not documented, it seems clear that current LLMs are nowhere near extrapolating to the outer boundary of all human knowledge (see Figure 1 again). They have no mechanism to incorporate what has never been written. In high-stakes domains like strategy, this means that LLMs can be powerful aides—generating drafts, brainstorming known solutions, and summarizing background research—yet they cannot replace the seasoned strategist's intuition or the creative leap that comes from real-world experience and tacit expertise.

8. Conclusion: Strategic Insight in the Age of Mean Articulation

The evidence so far suggests that LLMs remain more or less bounded by the data that shaped them but with masterful expertise at deploying multitudes of recombinations of that data in linguistically familiar and meaningful patterns. They are phenomenal at interpolation within their training set, but any true extrapolation into genuinely new knowledge or undocumented territory is, at best, highly constrained. The continuum of skills presented here serves as a map of the current landscape; it shows where LLMs shine, where they fall short, and by implication, where the frontier of their capabilities lies. This framework is intended to be one that researchers can expand, refine, and edit as LLM technology progresses.

For strategy scholars and practitioners, the lesson is not to dismiss LLMs—they (and their progeny" are here to stay—but to appropriate them relative to the task needed given their usefulness at that task. Used at the left end, they provide enormous productivity gains: drafting documents, synthesizing literatures, analyzing standard options, generating baseline strategies, or doing the significant grunt work that supports any genuinely creative insight. Used at the right end, however, they risk reinforcing herd behavior, masking anomalies, and failing to generate the rare insights that drive competitive advantage. In short, LLMs are powerful knowledge stores for humanity's expansive base of documented knowledge. They extend our reach into consensus knowledge across the majority of strategic cognition tasks but falter where contrarian vision is required. Recognizing the continuum of task difficulty could help organizations harness their strengths while guarding against misplaced reliance. Strategic breakthroughs for now will continue to demand the uniquely human capacity for genuinely novel imaginative leaps, causal reasoning, and daring departures from consensus, not to mention all of the undocumented implicit knowledge, intuitions, and skills forged in the fires of the real world. LLMs

can map much of the terrain of what is already known; for now, only human judgment can do significant extrapolation and chart the paths beyond it.

Going forward, future research will probe and push the limits of LLM interpolation. A growing number of leading researchers (e.g., Wooldridge 2020) argue that progress in LLMs will require new architectures or objectives beyond the current Transformer-based, next-token prediction LLM paradigm. LeCun (2022a, b) has been especially vocal in promoting alternative architectures (like "JEPAs"). Rather than reconstructing surface tokens, these approaches train models to predict in "embedding space," encouraging them to learn more abstract representations of the world and supporting capabilities, such as reasoning and planning (Dawid and LeCun 2023, Huang et al. 2025). Bengio (2024) has likewise advanced proposals that use "Transformers with Independent Mechanisms," which partition computation into modular subcomponents that can specialize and recombine flexibly (Lamb et al. 2021, Poli et al. 2023). These approaches aim to address some of the fundamental weaknesses of current LLM mean articulation architecture, including poor generalization to out-of-distribution problems, lack of causal reasoning, and inefficiency at long-context modeling. If they work, we can bet that AI firms will use them soon.

The conclusion here tempers the more extravagant hopes for the current architecture of LLMs in strategic decision making, but hopefully it amplifies and clarifies the domain of their utility. Human expertise and judgment remain crucial in domains that involve uncertainty, creativity, and tasks on the far right of the continuum. We might discover new technical approaches to extend the current limits of LLMs, or we might validate the indispensability of human cognition in areas beyond the scope of what large language models can currently achieve. Either outcome will deepen our understanding of intelligence, both artificial and human, and help us better harness these powerful mean articulation machines in concert with human skills to solve the challenges ahead.

Acknowledgments

The author thanks Jeff Yoshimi, Bob Horn, and the rest of the "Can Computers Think?" maps research team, as well as the organizers and attendees of the "Can AI do Strategy?" Sundance conference where most of this manuscript came together. In memoriam John Searle (1932–2025).

Endnotes

[1] "AGI" is the hypothetical form of artificial intelligence that can understand, learn, and perform *any* intellectual task that a human can across all domains without being restricted to narrow, specialized functions (Goertzel 2007). "Strong AI" is the hypothetical form of AI that successfully instantiates (rather than merely replicates) actual human intelligence (Searle 1980).

[2] Unfortunately, LLMs are not sharing their linguistic secrets here because these patterns, like all LLM "knowledge," are effectively a black box.

[3] Interpolation is the process of estimating or generating values *within* the range spanned by observed data points. Extrapolation is the process of estimating or generating values outside the range of observed data points.

[4] The Minsky (1974) "frame problem" is about how to create an artificial system that for any given action, can specify all of the things that *do not* change. I see the frame problem as a more specific variant of the larger problem of relevance—understanding what is relevant to any situation or expression, a problem that most believed would require revolutionary breakthroughs in algorithms and deep understanding of the workings of the actual human mind.

[5] One challenge with LLM assessment is that the most publicized failures are often manually patched by the AI firm. This is the case with the Wason Selection Task, the "reversal curse" where an LLM failed to understand that Alice is Bob's daughter after being told that Bob is Alice's father. This is also why some, like François Chollet, maintain a secret stock of LLM failures cases, which he uses to test new LLM releases.

References

Adner R, Levinthal DA (2008) Doing versus seeing: Acts of exploitation and perceptions of exploration. *Strategic Entrepreneurship J.* 2(1):43–52.

Ansoff HI (1965) *Corporate Strategy: An Analytic Approach to Business Policy for Growth and Expansion* (McGraw-Hill, Columbus, OH).

Apple Machine Learning Research Team (2025) *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity* (Apple Inc., Cupertino, CA).

Asai M, Kajino H, Fukunaga A, Muise C (2022) Classical planning in deep latent space. *J. Artificial Intelligence Res.* 74:1599–1686.

Bai Y, Chen F, Wang H, Xiong C, Mei S (2023) Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Adv. Neural Inform. Processing Systems* 36:2498.

Bai Y, Tu S, Zhang J, Peng H, Wang X, Lv X, Li J (2025) Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *Proc. 63rd Annual Meeting Assoc. Comput. Linguistics,* Long Papers, vol. 1 (Association for Computational Linguistics, Cedarville, OH), 3639–3664.

Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Drain D, et al. (2022) Training a helpful and harmless assistant with reinforcement learning from human feedback. Preprint, submitted April 12, https://arxiv.org/abs/2204.05862.

Bamberger PA (2018) AMD—Clarifying what we are about and where we are going. *Acad. Management Discoveries* 4(1):1–10.

Barney JB (1991) Firm resources and sustained competitive advantage. *J. Management* 17(1):99–120.

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? *Proc. 2021 ACM Conf. Fairness, Accountability, Transparency (FAccT '21)* (ACM, New York), 610–623.

Bengio Y (2024) Machine learning and information theory concepts towards a theory of mathematical intelligence. Preprint, submitted March 7, https://arxiv.org/abs/2403.04571.

Berg JM, Duguid MM, Goncalo JA, Harrison SH, Miron-Spektor E (2023) Escaping irony: Making research on creativity in organizations more creative. *Organ. Behav. Human Decision Processes* 175:104235.

Bhagavatula C, Le Bras R, Malaviya C, Sakaguchi K, Holtzman A, Rashkin H, Downey D, Yih WT, Choi Y (2020) Abductive commonsense reasoning. *8th Internat. Conf. Learn. Representations (ICLR 2020)* (OpenReview).

Bhardwaj A, Sergeeva A, Mahoney J, Nickerson J (2025) Theorizing as problem solving: A pragmatist perspective on the logic of pursuit. *Strategy Sci.* 10(4):338–359.

Bi J (2025) Don't believe AI hype, this is where it's actually headed | Oxford's Michael Wooldridge | AI history. *YouTube* (March 11), https://www.youtube.com/watch?v=Zf-T3XdD9Z8.

Binz M, Schulz E (2023) Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. USA* 120(6):e2218523120.

Blohm I, Antretter T, Sirén C, Grichnik D, Wincent J (2022) It's a people's game, isn't it?! A comparison between the investment returns of business angels and machine learning algorithms. *Entrepreneurship Theory Practice* 46(4):1054–1091.

Boden MA (2004) *The Creative Mind: Myths and Mechanisms*, 2nd ed. (Routledge, London).

Bommarito MJ II, Katz DM (2022) GPT takes the bar exam. Preprint, submitted December 29, https://doi.org/10.48550/arXiv.2212.14402.

Bridgelall R (2024) Unraveling the mysteries of AI chatbots. *Artificial Intelligence Rev.* 57:89.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, et al. (2020) Language models are few-shot learners. *Adv. Neural Inform. Processing Systems* 33:159.

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, et al. (2023) Sparks of artificial general intelligence: Early experiments with GPT-4. Preprint, submitted April 13, https://arxiv.org/abs/2303.12712.

Burgelman RA (1983) A process model of internal corporate venturing in the diversified major firm. *Admin. Sci. Quart.* 28(2):223–244.

Cai K (2023) Databricks CEO Ali Ghodsi's AI obsession made him a billionaire. *Forbes* (March 29), https://www.forbes.com/sites/kenrickcai/2023/03/29/databricks-ceo-ali-ghodsi-ai-obsession-billionaire/.

Cai K, Singh J (2025) Google clinches milestone gold at global math competition, while OpenAI also claims win. *Reuters* (July 22), https://www.reuters.com/world/asia-pacific/google-clinches-milestone-gold-global-math-competition-while-openai-also-claims-2025-07-22/.

Carlini N, Tramèr F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown TB, Song D, Erlingsson Ú, Oprea A, Raffel C (2021) Extracting training data from large language models. Bailey M, Greenstadt R, eds. *Proc. 30th USENIX Security Sympos. (USENIX Security '21)* (USENIX Association, Berkely), 2633–2650.

Chalmers D, MacKenzie NG, Carter S (2021) Artificial intelligence and entrepreneurship: Implications for venture creation in the fourth industrial revolution. *Entrepreneurship Theory Practice* 45(5):1028–1053.

Csaszar FA, Ketkar H, Kim H (2024) Artificial intelligence and strategic decision-making: Evidence from entrepreneurs and investors. *Strategy Sci.* 9(4):322–345.

Cyert RM, March JG (1963) *A Behavioral Theory of the Firm* (Prentice Hall, Hoboken, NJ).

Dawid A, LeCun Y (2023) Introduction to latent variable energy-based models: A path towards autonomous machine intelligence. Preprint, submitted June 5, https://arxiv.org/abs/2306.02572.

Dierickx I, Cool K (1989) Asset stock accumulation and sustainability of competitive advantage. *Management Sci.* 35(12):1504–1511.

Dunne D, Martin R (2006) Design thinking and how it will change management education: An interview and discussion. *Acad. Management Learn. Ed.* 5(4):512–523.

Einstein A (1952) *Relativity: The Special and the General Theory*, 15th ed. (Crown Publishers, New York).

Eisenhardt KM, Martin JA (2000) Dynamic capabilities: What are they? *Strategic Management J.* 21(10–11):1105–1121.

Fadell T (2012) The birth of the iPod and the iPhone. McAfee A, Brynjolfsson E, eds. *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy* (Harvard Business Review Press, Brighton, MA), 117–125.

Farrell H, Gopnik A, Shalizi C, Evans J (2025) Large AI models are cultural and social technologies. *Science* 387(6739):1153–1156.

Felin T, Holweg M (2024) Theory is all you need: AI, human cognition, and causal reasoning. *Strategy Sci.* 9(4):346–371.

Felin T, Zenger TR (2017) The theory-based view: Economic actors as theorists. *Strategy Sci.* 2(4):258–271.

Felin T, Gambardella A, Stern S, Zenger TR (2020) Lean startup and the business model: Experimentation revisited. *Long Range Planning* 53(4):101889.

Forrester JW (1961) *Industrial Dynamics* (MIT Press, Cambridge, MA).

Foss NJ, Klein PG (2012) *Organizing Entrepreneurial Judgment: A New Approach to the Firm* (Cambridge University Press, Cambridge, UK).

Franceschelli G, Musolesi M (2025) On the creativity of large language models. *AI Soc.* 40:3785–3795.

Freeman RE (1984) *Strategic Management: A Stakeholder Approach* (Pitman, Lanham, MD).

Frieder S, Pinchetti L, Chevalier A, Griffiths R-R, Salvatori T, Lukasiewicz T, Petersen PC, Berner J (2023) Mathematical capabilities of ChatGPT. Preprint, submitted July 20, https://arxiv.org/abs/2301.13867.

Gary MS, Wood RE (2011) Mental models, decision rules, and performance heterogeneity. *Strategic Management J.* 32(6):569–594.

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 (2005) Cognition and hierarchy: Rethinking the microfoundations of capabilities' development. *Organ. Sci.* 16(6):599–617.

Gavetti G (2012) Toward a behavioral theory of strategy. *Organ. Sci.* 23(1):267–285.

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

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

Gervás P (2024) Computational creativity: A critical perspective. *AI Magazine* 45(1):12–25.

Ghemawat P (2002) Competition and business strategy in historical perspective. *Bus. History Rev.* 76(1):37–74.

Goertzel B (2007) Human-level artificial general intelligence and the possibility of a technological singularity. *Artificial Intelligence* 171(18):1161–1173.

Greve HR (2003) *Organizational Learning from Performance Feedback: A Behavioral Perspective on Innovation and Change* (Cambridge University Press, Cambridge, UK).

Hamel G, Prahalad CK (1994) *Competing for the Future* (Harvard Business School Press, Boston).

Hanna RE, Smith LR, Mhaskar R, Hanna K (2024) Performance of language models on the family medicine in-training exam. *Family Medicine* 56(9):555–560.

Hanson NR (1958) *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science* (Cambridge University Press, Cambridge, UK).

Hao S, Gu Y, Ma H, Hong JJ, Wang Z, Wang DZ, Hu Z (2023). Reasoning with language model is planning with world model. *Proc. 2023 Conf. Empirical Methods Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA), 8154–8173.

Hofstadter D (2023) Gödel, Escher, Bach, and AI. *Atlantic* (July 5), https://www.theatlantic.com/ideas/archive/2023/07/godel-escher-bach-geb-ai/674589/.

Holton G (1973) *Thematic Origins of Scientific Thought: Kepler to Einstein* (Harvard University Press, Cambridge, MA).

Holtzman A, Buys J, Du L, Forbes M, Choi Y (2020) The curious case of neural text degeneration. *Internat. Conf. Learn. Representations (ICLR 2020)* (OpenReview).

Hsieh C-P, Sun S, Kriman S, Acharya S, Rekesh D, Jia F, Zhang Y, Ginsburg B (2024) RULER: What's the real context size of your long-context language models? Preprint, submitted August 6, https://arxiv.org/abs/2404.06654.

Huang H, LeCun Y, Balestriero R (2025) LLM-JEPA: Large language models meet joint embedding predictive architectures. Preprint, submitted October 7, https://arxiv.org/abs/2509.14252.

Isaacson W (2011) *Steve Jobs* (Simon & Schuster, New York).

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Survey of hallucination in natural language generation. *ACM Comput. Surveys* 55(12):248.

Jia N, Luo X, Fang Z, Liao C (2024) When and how artificial intelligence augments employee creativity. *Acad. Management J.* 67(1): 5–32.

Kambhampati S (2024) Can large language models reason and plan? *Ann. New York Acad. Sci.* 1534(1):15–18.

Kambhampati S, Valmeekam K, Guan L, Verma M, Stechly K, Bhambri S, Saldyt L, Murthy A (2024) LLMs can't plan, but can help planning in LLM-modulo frameworks. Preprint, submitted June 12, https://arxiv.org/abs/2402.01817.

Kaplan S (2008) Framing contests: Strategy making under uncertainty. *Organ. Sci.* 19(5):729–752.

Kaplan RS, Norton DP (1996) *The Balanced Scorecard: Translating Strategy into Action* (Harvard Business School Press, Boston).

Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D (2020) Scaling laws for neural language models. Preprint, submitted January 23, https://arxiv.org/abs/2001.08361.

Ketokivi M, Mantere S (2010) Two strategies for inductive reasoning in organizational research. *Acad. Management Rev.* 35(2):315–333.

Kim WC, Mauborgne R (2005) *Blue Ocean Strategy: How to Create Uncontested Market Space and Make the Competition Irrelevant* (Harvard Business School Press, Boston).

Kirzner IM (1973) *Competition and Entrepreneurship* (University of Chicago Press, Chicago).

Knight FH (1921) *Risk, Uncertainty, and Profit* (Houghton Mifflin, Boston).

Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y (2022) Large language models are zero-shot reasoners. *Adv. Neural Inform. Processing Systems* 35:22199–22213.

Kurzweil R (2022) *The Singularity Is Nearer* (Viking, New York).

Laban P, Hayashi H, Zhou Y, Neville J (2025) LLMs get lost in multi-turn conversation. Preprint, submitted May 9, https://arxiv.org/abs/2505.06120.

Lamb A, He D, Goyal A, Ke G, Liao C-F, Ravanelli M, Bengio Y (2021) Transformers with competitive ensembles of independent mechanisms. Preprint, submitted February 27, https://arxiv.org/abs/2103.00336.

Lampinen AK, Dasgupta I, Chan SCY, Sheahan HR, Creswell A, Kumaran D, McClelland JL, Hill F (2024) Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus* 3(7):233.

LeCun Y (2022a) AI and the limits of language. *LinkedIn* (September 9), https://www.linkedin.com/posts/yann-lecun_ai-and-the-limits-of-language-activity-6967929903409205248–ypi.

LeCun Y (2022b) A path towards autonomous machine intelligence. Position paper, OpenReview.

Levinthal DA (1997) Adaptation on rugged landscapes. *Management Sci.* 43(7):934–950.

Lin S, Hilton J, Evans O (2022) TruthfulQA: Measuring how models mimic human falsehoods. *Proc. 60th Annual Meeting Assoc. Comput. Linguistics,* Long Papers, vol. 1 (Association for Computational Linguistics, Cedarville, OH), 3214–3252.

Lovallo D, Clarke C, Camerer C (2011) Robust analogizing and the outside view: Two empirical tests of case-based decision making. *Strategic Management J.* 33(5):496–512.

Makridakis S, Hogarth RM, Gaba A (2010) Why forecasts fail—And what to do instead. *MIT Sloan Management Rev.* 51(2):83–90.

March JG (1991) Exploration and exploitation in organizational learning. *Organ. Sci.* 2(1):71–87.

March JG, Simon HA (1958) *Organizations* (Wiley, New York).

Marcus G (2022) Deep learning is hitting a wall. *Nautilus* (March 10), https://nautil.us/deep-learning-is-hitting-a-wall-238440/.

Marcus G, Davis E (2020) *Rebooting AI: Building Artificial Intelligence We Can Trust* (Vintage, New York).

Martin RL (2009) *The Design of Business: Why Design Thinking Is the Next Competitive Advantage* (Harvard Business Press, Boston).

McBride R, Packard MD, Clark BB (2024) Rogue entrepreneurship. *Entrepreneurship Theory Practice* 48(1):392–417.

McCoy R, Yao S, Friedman D, Hardy M, Griffiths T (2023) Embers of autoregression: Understanding large language models through the problem they are trained to solve. Preprint, submitted September 24, https://arxiv.org/abs/2309.13638.

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophysica* 5(4):115–133.

Minsky M (1974) A framework for representing knowledge. MIT Artificial Intelligence Laboratory Memo No. 306, MIT, Cambridge, MA.

Mintzberg H (1994) *The Rise and Fall of Strategic Planning* (Free Press, New York).

Mintzberg H, Waters JA (1985) Of strategies, deliberate and emergent. *Strategic Management J.* 6(3):257–272.

Mitchell M (2021) *Artificial Intelligence: A Guide for Thinking Humans* (Picador, New York).

Mossberg W (2007) The iPhone matches rivals' features but its main strength may be ease of use. *Wall Street J.* (June 27), https://www.wsj.com/articles/SB118289311361649057?mod=hp_lead_pos2.

Nelson RR, Winter SG (1982) *An Evolutionary Theory of Economic Change* (Harvard University Press, Cambridge, MA).

Nonaka I, Takeuchi H (1995) *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation* (Oxford University Press, Oxford, UK).

OpenAI (2023) GPT-4 technical report. Preprint, submitted December 19, https://arxiv.org/abs/2303.08774v4.

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, et al. (2022) Training language models to follow instructions with human feedback. *Adv. Neural Inform. Processing Systems* 35:27730–27744.

Pais A (1982) *Subtle Is the Lord: The Science and the Life of Albert Einstein* (Oxford University Press, Oxford, UK).

Pearl J (2000) *Causality: Models, Reasoning, and Inference* (Cambridge University Press, Cambridge, UK).

Peirce CS (1955) *Philosophical Writings of Peirce* (Dover Publications, Garden City, New York).

Peteraf MA (1993) The cornerstones of competitive advantage: A resource-based view. *Strategic Management J.* 14(3):179–191.

Polanyi M (1966) *The Tacit Dimension* (University of Chicago Press, Chicago).

Poli M, Massaroli S, Nguyen EN, Fu DY, Dao T, Baccus S, Bengio Y, Ermon S, Ré C (2023) Hyena hierarchy: Towards larger convolutional language models. Preprint, submitted April 19, https://arxiv.org/abs/2302.10866.

Porter ME (1980) *Competitive Strategy: Techniques for Analyzing Industries and Competitors* (Free Press, New York).

Porter ME (1996) What is strategy? *Harvard Bus. Rev.* 74(6):61–78.

Porter ME (2008) The five competitive forces that shape strategy. *Harvard Bus. Rev.* 86(1):25–40.

Posen HE, Keil T, Kim S, Meissner F (2018) Renewing research on problemistic search: A review and research agenda. *Acad. Management Ann.* 12(1):208–251.

Powell TC, Lovallo D, Fox CR (2011) Behavioral strategy. *Strategic Management J.* 32(13):1369–1386.

Rahmandad H, Sterman JD (2012) Reporting guidelines for simulation-based research in the social sciences. *System Dynam. Rev.* 28(4):396–411.

Raisch S, Krakowski S (2021) Artificial intelligence and management: The automation–augmentation paradox. *Acad. Management Rev.* 46(1):192–210.

Ren R, Liu Y (2024) Towards understanding how transformers learn in-context through a representation learning lens. *Adv. Neural Inform. Processing Systems (NeurIPS 2024)*, vol. 37 (Neural Information Processing Systems Foundation, Inc., San Diego, CA).

Repenning NP (2002) A simulation-based approach to understanding the dynamics of innovation implementation. *Organ. Sci.* 13(2):109–127.

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088):533–536.

Rumelt RP (1991) How much does industry matter? *Strategic Management J.* 12(3):167–185.

Schendel DE, Hofer CW (1979) *Strategic Management: A New View of Business Policy and Planning* (Little, Brown, and Company, Boston).

Schoemaker PJH (1995) Scenario planning: A tool for strategic thinking. *Sloan Management Rev.* 36(2):25–40.

Schumpeter JA (1934) *The Theory of Economic Development* (Harvard University Press, Cambridge, MA).

Seals S, Shalin V (2024) Evaluating the deductive competence of large language models. *Proc. 2024 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technol. (NAACL 2024)*, Long Papers, vol. 1 (Association for Computational Linguistics, Cedarville, OH), 8614–8630.

Searle JR (1980) Minds, brains, and programs. *Behav. Brain Sci.* 3(3):417–424.

Searle JR (1983) *Intentionality: An Essay in the Philosophy of Mind* (Cambridge University Press, Cambridge, UK).

Searle JR (1990) Is the brain's mind a computer program? *Sci. Amer.* 262(1):26–31.

Senge PM (1990) *The Fifth Discipline: The Art and Practice of the Learning Organization* (Doubleday/Currency, New York).

Sergeeva A, Bhardwaj A, Dimov D (2021) In the heat of the game: Analogical abduction in a pragmatist account of entrepreneurial reasoning. *J. Bus. Venturing* 36(6):106158.

Shepherd DA, Majchrzak A (2022) Machines augmenting entrepreneurs: Opportunities (and threats) at the nexus of artificial intelligence and entrepreneurship. *J. Bus. Venturing* 37(5): 106227.

Shoemaker PJH (1996) Scenario planning: A tool for strategic thinking. *Sloan Management Rev.* 36(2):25–40.

Simon HA (1957) *Models of Man: Social and Rational* (John Wiley & Sons, Hoboken, NJ).

Solow RM (1987) We'd better watch out. [Review of Manufacturing Matters: The Myth of the Post-Industrial Economy, by S. S. Cohen & J. Zysman]. *The New York Times Book Rev.* 36 (July 12).

Sterman JD (1989) Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Sci.* 35(3):321–339.

Sterman JD (2000) *Business Dynamics: Systems Thinking and Modeling for a Complex World* (Irwin/McGraw-Hill, Columbus, OH).

Surowiecki J (2008) *The Wisdom of Crowds* (Anchor Books, New York).

Sutton R (2019) The bitter lesson. *Incomplete Ideas* (March 13), incompleteideas.net/IncIdeas/BitterLesson.html.

Tang Z, Kejriwal M (2024) Human-like cognitive patterns as emergent phenomena in LLMs. Preprint, submitted December 20, https://arxiv.org/abs/2412.15501.

Teece DJ, Pisano G, Shuen A (1997) Dynamic capabilities and strategic management. *Strategic Management J.* 18(7):509–533.

Thompson C (2012) *Smarter Than You Think: How Technology Is Changing Our Minds for the Better* (Penguin Press, New York).

Togelius J, Yannakakis GN (2024) Choose your weapon: Survival strategies for depressed AI academics. *Proc. IEEE* 112(1):4–11.

Uehara S (2025) Is scaling the key to an AI future? The "scaling hypothesis," alternative pathways, and the "narrative" we choose to believe. Marubeni Washington Report No. 20, Marubeni Institute.

Valmeekam K, Stechly K, Kambhampati S (2024) LLMs still can't plan; can LRMs? A preliminary evaluation of OpenAI's o1 on PlanBench. Preprint, submitted September 20, https://arxiv.org/abs/2409.13373.

Valmeekam K, Marquez M, Sreedharan S, Kambhampati S (2023) On the planning abilities of large language models: A critical investigation. Oh A, Neumann T, Globerson A, Saenko K, Hardt M, Levine S, eds. *Advances in Neural Information Processing Systems*, vol. 36 (Curran Associates, Inc., Red Hook, NY), 75993–76005.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. *Adv. Neural Inform. Processing Systems* 30:5998–6008.

Vogelstein F (2013) *Dogfight: How Apple and Google Went to War and Started a Revolution* (Farrar, Straus and Giroux, New York).

Von Oswald J, Niklasson E, Randazzo E, Sacramento J, Mordvintsev A, Zhmoginov A, Vladymyrov M (2023) Transformers learn in-context by gradient descent. Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, eds., *Proc. 40th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 202 (PMLR, New York), 35151–35174.

Wang X, Wei J, Schuurmans D, Le QV, Chi E, Narang S, Chowdhery A, Zhou D (2023) Self-consistency improves chain-of-thought reasoning in language models. *Internat. Conf. Learn. Representations (ICLR 2023)* (OpenReview).

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

Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models. Preprint, submitted October 10, https://arxiv.org/abs/2201.11903v5.

Weick KE (1989) Theory construction as disciplined imagination. *Acad. Management Rev.* 14(4):516–531.

Welleck S, Kulikov I, Roller S, Dinan E, Cho K, Weston J (2020) Neural text generation with unlikelihood training. *Internat. Conf. Learn. Representations (ICLR 2020)* (Open Review).

Wenger E, Kenett YN (2025) We're different, we're the same: Creative homogeneity across LLMs. Preprint, submitted January 31, https://arxiv.org/abs/2501.19361.

Wernerfelt B (1984) A resource-based view of the firm. *Strategic Management J.* 5(2):171–180.

Whittaker ET (1951) *A History of the Theories of Aether and Electricity*, vol. 2 (Thomas Nelson, Nashville, TN).

Wooldridge M (2018) *A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going* (Flatiron Books, New York).

Wooldridge M (2020) *A Brief History of Artificial Intelligence* (Flatiron Books, New York).

Wu D, Yang J, Wang K (2024) Exploring the reversal curse and other deductive challenges in large language models. *Patterns* 5(10):100945.

Xie SM, Raghunathan A, Liang P, Ma T (2022) In-context learning as implicit Bayesian inference. *Internat. Conf. Learn. Representations (ICLR 2022)* (OpenReview).

Yang Z, Du X, Li J, Zheng J, Poria S, Cambria E (2024) Large language models for automated open-domain scientific hypotheses discovery. *Findings of the Association for Computational Linguistics: ACL 2024* (Association for Computational Linguistics, Stroudsburg, PA), 13545–13565.

Yang Z, Liu W, Gao B, Xie T, Li Y, Ouyang W, Poria S, Cambria E, Zhou D (2024) MOOSE-Chem: Large language models for rediscovering unseen chemistry scientific hypotheses. Preprint, submitted October 28, https://arxiv.org/abs/2410.07076v3.

Zečević M, Willig M, Dhami DS, Kersting K (2023) Causal parrots: Large language models may talk causality but are not causal. Preprint, submitted August 23, https://arxiv.org/abs/2308.13067.

Zhang W, Zhang R, Guo J, de Rijke M, Fan Y, Cheng X (2024) Pretraining data detection for large language models: A divergence-based calibration method. *Proc. 2024 Conf. Empirical Methods Natural Language Processing (EMNLP 2024)* (Association for Computational Linguistics, Stroudsburg, PA), 5263–5274.

Zhou J, Sinha M, Dong J, Li Y, Hu Y (2024a) Enhancing logical reasoning in large language models via extract-then-answer prompting. Preprint, submitted December 16, https://doi.org/10.48550/arXiv.2409.12437.

Zhou J, Ghaddar A, Zhang G, Ma L, Hu Y, Pal S, Coates M, Wang B, Zhang Y, Hao J (2024b) Enhancing logical reasoning in large language models through graph-based synthetic data. Preprint, submitted December 16, https://arxiv.org/abs/2409.12437.

Zhou Y, Ye J, Ling Z, Han Y, Huang Y, Zhuang H, Liang Z, et al. (2025) Dissecting logical reasoning in large language models: A fine-grained evaluation and supervision study. Preprint, submitted October 9, https://arxiv.org/abs/2506.04810.

Russ McBride is an associate professor at UC Merced in the Department of the Management of Complex Systems. He received his master's from Stanford in symbolic systems before doing artificial intelligence (AI) research at the Lexington Institute, which produced the "Can Computers Think?" argumentation maps. He did his PhD at UC Berkeley and is now the Director of the Social Reality & Cognition Research Group (http://sorac.info), and currently teaches an "AI for Entrepreneurs" course, among others. His forthcoming theory is "Pragmatic Entrepreneurship."