Mean Articulation Machines

Russ McBride
University of California, Merced
russ.mcbride@ucmerced.edu

Abstract

The performance of 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 which 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 due to 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.

Keywords: A.I., Extreme Transformational Creativity, Large Language Models, Combinatorial Creativity, Decision Making, Entrepreneurship, Strategy Formulation, Strategy Implementation, Transformational Creativity.

1. Introduction

The long AI winters of previous decades now seem like distant history, and there is renewed optimism around deep neural network, large language models (LLMs). Enthusiasm for the prospects of AI is back in full bloom as is, apparently, an 'AI spring'. Nobel-laureate Robert Solow's (1987) quip that, "You 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 exam, and pass the Multistate Performance Law Test (Bommarito & Katz, 2022; Hanna, Smith, Mhaskar, & Hanna, 2024). But LLMs have not achieved much-hyped 'artificial general intelligence' (AGI) or 'Strong AI'.[1] And 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 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, e.g., the most amazing, articulate summary of, e.g., some strategic framework, just before failing a child's simple letter-shifting game (McCoy, Yao, Friedman, Hardy, Griffiths, 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 re-deployment machines', but that's a mouthful. For a simpler (and less accurate) shorthand, I will highlight this

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).

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. And 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, then? 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 & Waters, 1985; Shoemaker, 1996; Powell, Lovallo, & Fox, 2011; Gavetti, 2012). A continuum of increasingly difficult strategic cognition skills will be explored which will illustrate how far along that continuum 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—a 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 can not 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 super-weakness.

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 & Pitts, 1949), 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 Rumelhart, Hinton, & Williams's (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 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 multi-head 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 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 anti-climatic and depressing to some life-long AI researchers, like Sutton (2019), who famously talked about the "bitter lesson" from this, or Hofstadter (2023), or Togelius & 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. Though 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 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's 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: If human knowledge leads inevitably to further advances in human knowledge, so too should the proxy structures of human

2 Unfortunately, LLMs aren't sharing their linguistic secrets here because these patterns, like all LLM 'knowledge' is 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.

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, "How far can LLM models extrapolate beyond the text and text patterns upon which they are trained?" (See Figure 1.)

[INSERT FIGURE 1 ABOUT HERE]

Optimism drives grand ambitions that LLMs (or some kind of machine learning system) can, first, accurately capture all the knowledge stored in the vast corpora of documented knowledge, then second, extend to all current *undocumented* human knowledge, and then third, perhaps lead to advances in science, humanities, and civilization-altering genuine knowledge breakthroughs. How far could such patterns carry us? All 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, "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 it hides between the lines? Undocumented knowledge? Future breakthroughs?

The ultimate extent to which an LLM can extrapolate beyond its training data is currently unknown and heavily debated. LLM optimists (e.g., Kurtzweil, 2022; Bubeck et al., 2023; Kaplan et al., 2020; 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 Marcus (2022), LeCun (2022), Wooldridge (2020), and Bender et al. (2021), 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 LLMs extrapolation boundary is the ongoing goal of LLM engineers. Live research questions are: "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?"; and, in general, "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 unbridgeable chasm between syntax and semantics, most clearly described in Searle's "Chinese Room Argument" (1980, 1990), 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 the ones 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 (Bridgelall, 2024; Zhang et al., 2024; McCoy et al., 2023; Marcus, 2022; Bender et al., 2021). The result is a machine that contains the full distribution of the training data, but 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, while 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, Xu, Liu, & Sun (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, since 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 (Bridgelall, 2024; Zhang et al., 2024; Marcus, 2022; Bender et al., 2021; McCoy et al., 2023).

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

4 Minsky's (1974) 'frame problem' is about how to create an artificial system that, for any given action, can specify all the things that *don't* 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.

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, e.g., 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 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 don't 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 that 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 the restatement of widely known concepts. It is both this superpower of LLMs, which is simultaneously its kryptonite, that explains the hallucinations, and the strange failures we see, like the inability to solve a problem when reworded in a unfamiliar way. Wooldridge, discussing his 2018 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 versus 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 enables 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 super-powered 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: First, it's 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 Appendix A and some suggested learning material in 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

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 they do, toward the goal of understanding their performance in strategy-specific tasks. I won't 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 sub-skills 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 can individuate a concept from syntactically disparate expressions of it.

Search

Search, in strategy, is the process of generating a broad range of strategic alternatives, solutions, or options for the firm (Cyert & March, 1963; Ansoff, 1965; Nelson & Winter, 1982; Levinthal, 1997; March, 1991; 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 the LLM has been exposed to in training data or relatively simple recombinations thereof. There are, it should be noted though, 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—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.

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 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.

[INSERT FIGURE 2 ABOUT HERE]

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 capabilities of LLMs, as does formatting a response to a document template (e.g., introduction, methods, results, conclusion), or imitating the writing style of Wordworth 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'.

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 (Ouyang et al., 2022; Askell et al., 2021; Bai et al., 2022). This often requires more reinforcement training from actual human feedback (RLHF) 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 prompting and unlikelihood training) (Wei et al., 2022; Welleck et al., 2019; Holtzman et al., 2020).

Strategic Framework Assessment

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

Analogical reasoning, combinatorial creativity with familiar components, simple strategy ideation, high-frequency information with a conflict, induction, math

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

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 & 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 startup's financial platform—leveraging insights from a seemingly unrelated context to solve a strategic problem at hand. (Gary, Wood, & Pillinger, 2012; Lovallo, Clarke, & Camerer, 2011). 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, Holyoak, & Lu, 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).

Combinatorial creativity with familiar components

Boden's (2004) distinction between simple "combinatorial creativity" versus novel "transformational creativity" (to which I will add the category of "extreme transformational creativity" later) is useful here. LLMs excel at simple combinatorial creativity, mixing ideas and patterns they've seen, but stumble at truly novel creativity, especially when it violates existing, established knowledge in the extreme cases (Gervás, 2024; Franceschelli, & 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.

[INSERT FIGURE 3 ABOUT HERE]

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 Boden’s (2004) ‘combinatorial creativity’. 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 humans judge as creative (in this sense) when outputs involve rearrangements of well-represented concepts (Binz & 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 & Kenett, 2025). LLMs remix existing knowledge into new but statistically coherent forms.

Simple strategy ideation

This requires combinatorial creativity and overlaps with aggregation, so 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 mid-priced home fitness equipment company”, and get reasonable suggestions like: adding a loyalty program or bundling extended warranties.

High-frequency ideas that conflict with some information

High-frequency ideas in training data is much more common than contradictory information. When an LLM encounters debates—e.g., Copenhagen vs. 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 states “Shakespeare wrote Hamlet” and 10% attributes it to Marlowe, default decoding will almost always return “Shakespeare,” since low-probability continuations are suppressed (Bridgelall, 2024; Zhang et al., 2024; Holtzman et al., 2020). 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, Hilton, & Evans, 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 reduces but does not eliminate this "consensus bias," and later evaluations (OpenAI, 2023; Ji et al., 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.

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 & Simon, 1958; March, 1991; Gavetti & Levinthal, 2000; 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 & 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; Xie et al., 2022; von Oswald 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 & Schulz, 2023; Ren et al., 2024, McCoy et al., 2023). 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.

Math

Although mathematics is formally a deductive activity (usually), its role in strategic cognition is largely practical: strategists routinely need to compute, e.g., breakeven 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 (Makridakis et al., 2010; Kaplan & Norton, 1996). Modern LLMs handle such routine, well-specified computations effectively, and recent systems have even reached gold-medal–level performance on structured Olympiad-style tasks under heavy scaffolding (Cai & 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 (Frieder et al., 2023; Wang 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 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 vs. 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 (Mintzberg, 1994; Hamel & Prahalad, 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

Abductive reasoning, deductive reasoning, causal reasoning, novel idea generation with minimal conflict, 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's often because they've seen the problem (or a similar pattern) in their training.

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 (Hanson, 1958; Peirce, 1955). Abduction allows strategists to leap from incomplete or ambiguous data to a coherent explanation (Dunne & Martin, 2006; Martin, 2009; Sergeeva, Bhardwaj, & Dimov, 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 (Ketokivi & Mantere, 2010; Weick, 1989). 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 (Bamberger, 2018; Nonaka & Takeuchi, 1995). 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 don't match prior exposure. If all the LLM training data showed that "the butler did it" and there's no

butler mentioned in the IBM case, then it will get confused. This is a controversial position though, with Berg, et al. (2023) and Wei et al. (2022) suggesting the just opposite.

Bhagavatula et al. (2020) introduced the first "abductive commonsense 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) were an explanation for the observations had to be generated from scratch. 2020 models performed at about a 45% level compared to 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 don't yet have a model that can best human performance here.

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 cannot 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 (Teece et al., 1997; Levinthal, 1997; Gavetti & Levinthal, 2000).

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 CoT (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 multi-step reasoning tasks when prompted to generate explicit intermediate reasoning steps (Wei et al.,

2022; Kojima 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., 2023). In mathematics and deductive reasoning, CoT substantially improves accuracy by enforcing stepwise constraint satisfaction and reducing omitted or inconsistent intermediate steps (Wei et al., 2022; Wang et al., 2023). 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, Fu, Liu, Xie, Lin, Wang, Cai, & Wu, 2024). 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 et al., 2024). Chain-of-thought improves presentation but not underlying validity (Tang et al., 2024; Zhou et al., 2025). Apple's ML 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 eleven 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 13—LLMs succeed only if the specific variant is common in training. They failed on rot-2 or rot-8, despite the task being simple and structurally identically to the more frequently-seen rot-13 (pp. 13–15). Replacing each letter with one 2 letter ahead or 8 or 13 shouldn't make any difference; it's 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 a famous and frequently mentioned figures (e.g., Jeff Bezos) versus ~23% for infrequent ones (pp. 31–32). Thus, LLM output quality tracks mention frequency reflecting the frequency bias inherent in the architecture. For strategists, this means 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.

[INSERT FIGURE 4 ABOUT HERE]

Strategy frequently requires inference from sparse, unusual, or novel data—precisely where LLMs degrade. Even small syntactic changes ("12 × 3" vs. "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 et al., 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, DeepSeek-R1) underperform humans on multi-step, 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

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 LLMs 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.

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 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).

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 & 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; Power & Reid, 2005; 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, 2020). 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 (Felin & Holweg, 2024; Rahmandad & Sterman, 2012). 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 & Davis, 2020, p. 112, also: Zečević, Willig, Dhami & Kersting, 2023; Zhou et al., 2024; Wu et al., 2024), 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 2020–2025 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 fail when causal structure must be inferred, when background knowledge is sparse, or when reasoning involves complex multi-variable 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.

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, Xu, Liu, Li, & Sun, 2024). Truly novel idea are then mostly hopeless for LLMs and any form of conflicting information reduces their chances for success even further. This is Boden's (2004) 'transformational creativity' which requires "changing the rules of the conceptual space, so that ideas can be generated which were impossible before" (Boden, 2004, p6). 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.

[INSERT FIGURE 5 ABOUT HERE]

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 & Mauborgne, 2005). As such, this is not a simple recombination of existing ideas. Sustained advantage typically requires a departure from standard ideation, since 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 Boden's (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 since 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 further right on the task continuum.

The architecture LLMs is mismatched to this task. They are biased toward high-frequency text continuations and away from rare or counter-intuitive propositions (Bender et al., 2021; Zhang et al., 2024). Empirical work shows that when truth conflicts with high-frequency beliefs, models tend to reproduce the latter (Lin, Hilton, & Evans, 2022). Creativity studies find that LLM-generated ideas are appropriate recombinations of common elements but converge toward a narrow set of themes and avoid genuinely unusual structures (Binz & Schulz, 2023; Wenger & 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, MacKenzie, & Carter, 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). 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 et al., 2024a; 2024b).

Accordingly, LLMs are useful for simple strategy ideation (further left on the continuum)—listing business-model variants, surfacing analogies, or recombining familiar strategic elements (Hamel &

Prahalad, 1994; Gavetti & Rivkin, 2005; Csaszar, Ketkar, & Kim, 2024). They extend exploitation and local search (March, 1991; Gavetti & Levinthal, 2000). But because they articulate documented ideas, they are unlikely to generate VRIO strategies, radical business-model innovations, or counter-consensus value theories (Barney, 1991; Felin & 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, PACE), entrepreneurship (e.g., new-product or opportunity-recognition studies), or scientific hypothesis generation (e.g., MOOSE, 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.

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.

Decision-making under uncertainty

Often, decisions must be made without complete information. Strategic decisions typically fall on Knight's "uncertainty" side of risk versus uncertainty (Knight, 1921) dichotomy where situations have unknowable probabilities because, e.g., 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, Lovallo, & Fox, 2011). Entrepreneurs similarly exercise what Kirzner (1973) called 'alertness': the ability to notice overlooked possibilities in environments where information is sparse or equivocal. And, 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 (Gavetti, 2012; Gary & Wood, 2011). 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 McCoy et al's work, 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, Hilton, & Evans, 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, non-obvious causal models, or low-frequency signals often matter most (Felin & Zenger, 2017; Gavetti, 2012). 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 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 counter-consensus 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.

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 multi-year 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, Hogarth, & Gaba, 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 (Gavetti & Levinthal, 2000; Teece, Pisano, & Shuen, 1997).

In computer science, however, long-range planning refers to the ability of an algorithm or agent to execute extended, multi-step 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, 2022; 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. 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. While more memory helps, models often ignore or mis-prioritize earlier steps (Hsieh et al., 2024; LongBench v2 Team, 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, 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, Rovers) to long-horizon robotics suites (Meta-World, CALVIN, ManiSkill2), exploration-heavy RL environments (Montezuma's Revenge, Pitfall), tool-use settings (WebArena, BrowserGym), and language-based planning evaluations (PlanBench, ALFWorld)—the empirical pattern is consistent: LLMs perform poorly on all tasks that require genuine long-horizon planning (Asai & Fukunaga 2022; Hao et al. 2023; Valmeekam et al., 2023; Valmeekam et al., 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%, with performance collapsing as horizon length grows. In robotics

benchmarks, state-of-the-art RL agents reach 60–90% success on long-horizon tasks, while LLM controllers remain in the 10–40% range. This is one area where progress has gone backward compared to "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 multi-step 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 multi-step dependencies.

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 doesn't 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 Boden's (2004) category. Of course, extreme transformational creativity is no easy task for humans, either, but 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's 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.

[INSERT FIGURE 6 ABOUT HERE]

There are no existing benchmarks—in computer science, psychology, economics, strategy, or creativity research—that measure the ability to generate truly novel, while still feasible, ideas which 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 further down the scale we move further 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, denoted by the dashed line, 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

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. Mobile design conventions held, first, 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 (Kahney, 2008). A comment from early 2007 captured the prevailing sentiment: “no 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, non-removable battery was almost unthinkable—a “jarring design choice” at the time (Kahney, 2008). Apple’s own engineers knew 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 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 due to CPU, memory, and power constraints. Industry experts assumed 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 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 multi-touch display, prioritizing optical clarity and scratch-resistance at the risk of fragility. After Jobs found 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 3G. By 2007, third-generation (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 a phone would omit a keyboard and stylus, use only a touchscreen with multi-touch 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 (Kahney, 2008). 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 candybar 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 counter-consensus ideas—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 & Zenger, 2017), if they are not 'counter prompted'. 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 (multi-touch 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. Barney's (1991) VRIO framework holds that a strategy yields sustained advantage only if it employs resources or decisions that are valuable, rare, inimitable, and non-substitutable. 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 (due to Apple's unique integration of software and hardware), and supported by the organization. It took 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 "the conventional wisdom might be wrong here." In the iPhone case, a human had to apply non-consensus reasoning—seeing, for example, that a multi-touch UI could work if done right, even though all prior evidence suggested 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 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.

Novel Theories

LLMs inability to compose genuinely novel ideas applies to generating genuinely novel theories as well. The theory-based view (Felin & 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, e.g., iterating a product idea from customer feedback, is essential for innovative advances (Felin, Gambardella, Stern, & Zenger, 2019). 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.

His 1905 development of special relativity provides another clear case of a task at the far end of the continuum. 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, Hilton, & Evans, 2022).

What made Einstein's insight radical was not a modest recombination of existing concepts but the rejection of multiple entrenched assumptions simultaneously. 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, 1905/1952). The weirdness of the theory is that, unlike all other objects whose speed varies with your speed (i.e., a car traveling at 50mph only appears to be traveling at 20mpg if you are traveling next to it at 30mph), the speed of light is absolute no matter how

fast your are traveling next to it. A language model trained on 1905-era discourse could 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 can not.

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 but 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 counter-consensus 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 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 counter-consensus 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, Hilton, & Evans, 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: 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 which 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, counter-cyclical 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 & 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 & Majchrzak, 2022).

The distinction between familiar tasks easy for LLMs and exploratory tasks that requires extrapolating beyond the training data echos March's (1991) distinction 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, the benefits of exploration include: innovation and novelty; adaptability in the form of building capacity to adapt to changing environments; learning under uncertainty; and 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's 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 one ever had that never requires sleep. The left side of the continuum highlights the immense value 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, and combinatorial creativity, and high-quality consensus summarization are well within their domain. For strategists, this means 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. We hope that other researchers might provide 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 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. And the broad answer is that the same feature which 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 they can be discovered through textual analysis. Alternative architectures custom-built for knowledge extension in 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?”; and, “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 interpolate 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 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 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. And 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”), are not only not documented but simply 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, 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.

[INSERT FIGURE 1 AGAIN ABOUT HERE]

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 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 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 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 & LeCun, 2023; Huang, LeCun, & Balestriero, 2025). Bengio and colleagues have likewise advanced proposals that use "Transformers with Independent Mechanisms (TIM)", which partition computation into modular sub-components that can specialize and recombine flexibly (Poli et al., 2023; Lamb et al., 2021). 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 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.

References

Adner, R., & Levinthal, D. A. (2008). Doing versus seeing: Acts of exploitation and perceptions of exploration. Strategic Entrepreneurship Journal, 2(1), 43–52. https://doi.org/10.1002/sej.19

Amabile, T. M. (2020). Creativity and the psychology of everyday genius. Harvard Business Review Press.

Amini, A., Gabriel, S., Lin, S., Koncel-Kedziorski, R., Choi, Y., & Hajishirzi, H. (2019). MathQA: Towards interpretable math word problem solving with operation-based formalisms. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 2357–2367). Association for Computational Linguistics.

Ansoff, H. I. (1965). Corporate strategy: An analytic approach to business policy for growth and expansion. McGraw-Hill.

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.

Asai, M., Kajino, H., Fukunaga, A., & Muise, C. (2022). Classical planning in deep latent space. Journal of Artificial Intelligence Research, 74, 1599–1686. https://doi.org/10.1613/jair.1.13768

Bai, Y., Chen, F., Wang, H., Xiong, C., & Mei, S. (2023). Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In Advances in Neural Information Processing Systems (Vol. 36).

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., … Amodei, D. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

Bamberger, P. A. (2018). AMD—Clarifying what we are about and where we are going. Academy of Management Discoveries, 4(1), 1–10. https://doi.org/10.5465/amd.2018.0003

Barney, J. B. (1986). Organizational culture: Can it be a source of sustained competitive advantage? Academy of Management Review, 11(3), 656–665.
Barney, J. B. (1991). Firm resources and sustained competitive advantage. Journal of Management, 17(1), 99–120.
Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). https://doi.org/10.1145/3442188.3445922
Bengio, Y. (2024). Machine learning and information theory concepts towards a theory of mathematical intelligence. arXiv. https://arxiv.org/abs/2403.04571
Berg, J. M., Duguid, M. M., Goncalo, J. A., Harrison, S. H., & Miron-Spektor, E. (2023). Escaping irony: Making research on creativity in organizations more creative. Organizational Behavior and Human Decision Processes, 175, 104235.
Bhagavatula, C., Le Bras, R., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, W.-t., & Choi, Y. (2020). Abductive commonsense reasoning. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020).
Bhardwaj, A., Sergeeva, A., Mahoney, J. T., & Nickerson, J. A. (2025). Theorizing as problem solving: A pragmatist perspective on the logic of pursuit. Strategy Science. Advance online publication. https://doi.org/10.1287/stsc.2023.0075
Bi, J. (2025, March 11). Don’t believe AI hype, this is where it’s actually headed | Oxford’s Michael Wooldridge | AI history [Video]. YouTube. https://www.youtube.com/watch?v=Zf-T3XdD9Z8
Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences, 120(6), e2218523120. https://doi.org/10.1073/pnas.2218523120
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 and Practice, 46(4), 1054–1091. https://doi.org/10.1177/1042258720945206
Bridgelall, R. (2024). Unraveling the mysteries of AI chatbots. Artificial Intelligence Review. https://doi.org/10.1007/s10462-024-10720-7
Boden, M. A. (2004). The creative mind: Myths and mechanisms (2nd ed.). Routledge. https://doi.org/10.4324/9780203508527
Bommarito, M. J., II, & Katz, D. M. (2022). GPT takes the bar exam. arXiv. https://doi.org/10.48550/arXiv.2212.14402
Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33).
Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv. https://arxiv.org/abs/2303.12712
Burgelman, R. A. (1983). A process model of internal corporate venturing in the diversified major firm. Administrative Science Quarterly, 28(2), 223–244.
Cai, K. (2023, March 29). Databricks CEO Ali Ghodsi’s AI obsession made him a billionaire. Forbes. https://www.forbes.com/sites/kenrickcai/2023/03/29/databricks-ceo-ali-ghodsi-ai-obsession-billionaire/
Cai, K., & Singh, J. (2025, July 22). Google clinches milestone gold at global math competition, while OpenAI also claims win. Reuters.
Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. In Proceedings of the 30th USENIX Security Symposium (USENIX Security 21) (pp. 2633–2650). USENIX Association.

Chalmers, D., MacKenzie, N. G., & Carter, S. (2021). Artificial intelligence and entrepreneurship: Implications for venture creation in the fourth industrial revolution. Entrepreneurship Theory and Practice, 45(5), 1028–1053. https://doi.org/10.1177/1042258720934581

Cheng, K., Ahmed, N. K., Willke, T. L., & Sun, Y. (2024). Structure guided prompt: Instructing large language model in multi-step reasoning by exploring graph structure of the text. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 9407–9430). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.528

Clark, A. (1997). Being there: Putting brain, body, and world together again. MIT Press.

Clausewitz, C. von. (1832/1976). On war. Princeton University Press.

Clark, P., Tafjord, O., & Richardson, K. (2020). Transformers as soft reasoners over language. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) (pp. 3882–3890). International Joint Conferences on Artificial Intelligence Organization.

Csaszar, F. A., Ketkar, H., & Kim, H. (2024). Artificial intelligence and strategic decision-making: Evidence from entrepreneurs and investors. Strategy Science, 9(4), 322–345.

Cyert, R. M., & March, J. G. (1963). A behavioral theory of the firm. Prentice Hall.

Dalvi, B., Jansen, P., Tafjord, O., Xie, Z., Smith, H., Pipatanangkura, L., & Clark, P. (2021). Explaining answers with entailment trees. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 7358–7370). Association for Computational Linguistics.

Dawid, A., & LeCun, Y. (2023). Introduction to latent variable energy-based models: A path towards autonomous machine intelligence. arXiv. arXiv:2306.02572.

Del, M., & Fishel, M. (2022). True detective: A deep abductive reasoning benchmark undoable for GPT-3 and challenging for GPT-4. arXiv. arXiv:2212.10114.

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

Dreyfus, H. L. (1972). What computers can’t do: A critique of artificial reason. Harper & Row.

Dreyfus, H. L. (1992). What computers still can’t do: A critique of artificial reason. MIT Press.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., & Gardner, M. (2019). DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. arXiv. arXiv:1903.00161.

Dunne, D., & Martin, R. (2006). Design thinking and how it will change management education: An interview and discussion. Academy of Management Learning & Education, 5(4), 512–523.

Dutton, J. E., & Jackson, S. E. (1987). Categorizing strategic issues: Links to organizational action. Academy of Management Review, 12(1), 76–90.

Einstein, A. (1952). *Relativity: The special and the general theory* (15th ed.). Crown Publishers. (Original work published 1905).

Eisenhardt, K. M., & Martin, J. A. (2000). Dynamic capabilities: What are they? Strategic Management Journal, 21(10–11), 1105–1121.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. OpenAI Whitepaper.

Fadell, T. (2012). The birth of the iPod and the iPhone. Harper Business.

Felin, T., Gambardella, A., Stern, S., & Zenger, T. R. (2019). Lean startup and the business model: Experimentation revisited. Long Range Planning, 53(4). https://doi.org/10.1016/j.lrp.2019.06.002

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

Felin, T., & Zenger, T. R. (2017). The theory-based view: Economic actors as theorists. Strategy Science, 2(4), 258–271. https://doi.org/10.1287/stsc.2017.0048

Fiol, C. M., & Huff, A. S. (1992). Maps for managers: Where are we? Where do we go from here? Journal of Management Studies, 29(3), 267–285. https://doi.org/10.1111/j.1467-6486.1992.tb00665.x.

Forrester, J. W. (1961). Industrial dynamics. MIT Press.

Foss, N. J., & Klein, P. G. (2012). Organizing entrepreneurial judgment: A new approach to the firm. Cambridge University Press.

Franceschelli, G., & Musolesi, M. (2025). On the creativity of large language models. AI & Society, 40, 3785–3795. https://doi.org/10.1007/s00146-024-02127-3
Freeman, R. E. (1984). Strategic management: A stakeholder approach. Pitman.
Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., & Berner, J. (2023). Mathematical capabilities of ChatGPT. arXiv. arXiv:2301.13867.
Gary, M. S., & Wood, R. E. (2011). Mental models, decision rules, and performance heterogeneity. Strategic Management Journal, 32(6), 569–594.
Gary, M. S., Wood, R. E., & Pillinger, T. (2012). Enhancing mental models, analogical transfer, and performance in strategic decision making. Strategic Management Journal, 33(11), 1229–1246. https://doi.org/10.1002/smj.1979.
Gavetti, G. (2000). Cognition and hierarchy: Rethinking the microfoundations of capabilities' development. Organization Science, 11(6), 599–617.
Gavetti, G. (2012). Toward a behavioral theory of strategy. Organization Science, 23(1), 267–285.
Gavetti, G., & Rivkin, J. W. (2005). How strategists really think: Tapping the power of analogy. Harvard Business Review, 83(4), 54–63.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 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. Business History Review, 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.
Grant, R. M. (1991). The resource-based theory of competitive advantage: Implications for strategy formulation. California Management Review, 33(3), 114–135.
Greve, H. R. (2003). Organizational learning from performance feedback: A behavioral perspective on innovation and change. Cambridge University Press.
Gundawar, A., Valmeekam, K., Verma, M., & Kambhampati, S. (2024). Robust planning with compound LLM architectures: An LLM-Modulo approach. arXiv. arXiv:2411.14484.
Guo, D., DeepSeek-AI, Liu, X., Zhang, H., Wang, L., Zhao, Y., … Tang, J. (2025). DeepSeek-R1: Incentivizing reasoning capability in large language models via reinforcement learning. Nature. Advance online publication. https://doi.org/10.1038/s41586-025-09422-z
Hamel, G., & Prahalad, C. K. (1994). Competing for the future. Harvard Business School Press.
Han, S., Schoelkopf, H., Zhao, Y., Qi, Z., Riddell, M., Zhou, W., Coady, J., Peng, D., Qiao, Y., Benson, L., Sun, L., Wardle-Solano, A., Szabó, H., Zhou, Y., Xiong, C., Ying, R., Cohan, A., & Radev, D. (2024). FOLIO: Natural language reasoning with first-order logic. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 22017–22031). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.1229
Hanna, R. E., Smith, L. R., Mhaskar, R., & Hanna, K. (2024). Performance of language models on the family medicine in-training exam. Family Medicine, 56(9), 555–560. https://doi.org/10.22454/FamMed.2024.233738 Journal of Medical Internet Research, 26(1), e66114.
Hanson, N. R. (1958). Patterns of discovery: An inquiry into the conceptual foundations of science. Cambridge University Press.
Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., & Hu, Z. (2023). Reasoning with language model is planning with world model. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 8154–8173). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.507
Hayward, M. L. A., Shepherd, D. A., & Griffin, D. (2006). A hubris theory of entrepreneurship. Management Science, 52(2), 160–172.
He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Hillebrand, L., Raisch, S., & Schad, J. (2025). Managing with artificial intelligence: An integrative framework. Academy of Management Annals, 19(1), 343–375.
Hodgkinson, G. P., & Healey, M. P. (2011). Psychological foundations of dynamic capabilities: Reflexion and reflection in strategic management. Strategic Management Journal, 32(13), 1500–1516.
Hofstadter, D. R. (2023, July 5). Gödel, Escher, Bach, and AI. The Atlantic.
Holton, G. (1973). Thematic origins of scientific thought: Kepler to Einstein. Harvard University Press.
Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In Proceedings of the International Conference on Learning Representations (ICLR 2020).
Horn, R., Yoshimi, J., Deering, M., & McBride, R. (1998). Can computers think: The history and status of the debate. MacroVu Inc.
Hsieh, C.-P., Chen, C.-Y., Chang, W.-C., et al. (2024). RULER: What's the real context size of your long-context LMs? arXiv. arXiv:2404.06654.
Huang, H., LeCun, Y., & Balestriero, R. (2025). LLM-JEPA: Large language models meet joint embedding predictive architectures. arXiv. arXiv:2509.14252.
Huang, Y., Liu, Y., Li, C., et al. (2024). CLOMO: Counterfactual logical modification with LLMs. In Findings of the Association for Computational Linguistics: ACL 2024.
Hume, D. (1739/1978). A treatise of human nature. Oxford University Press.
Isaacson, W. (2011). *Steve Jobs.* Simon & Schuster.
Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248: 1–38. https://doi.org/10.1145/3571730.
Jia, N., Luo, X., Fang, Z., & Liao, C. (2024). When and how artificial intelligence augments employee creativity. Academy of Management Journal, 67(1), 5–32.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
Kahney, L. (2008). Inside Steve's brain. Portfolio.
Kaplan, R. S., & Norton, D. P. (1996). *The balanced scorecard: Translating strategy into action*. Harvard Business School Press.
Kaplan, S. (2008). Framing contests: Strategy making under uncertainty. Organization Science, 19(5), 729–752.
Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., … Amodei, D. (2020). Scaling laws for neural language models. arXiv. arXiv:2001.08361.
Kadavath, S., Conerly, T., Askell, A., Henighan, T., Johnston, S., Kravec, S., Lin, C., Nadeau, M., Radhakrishnan, A., Tang, E., Tran-Johnson, E., Bowman, S. R., Chen, A., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Liang, P., Manning, C. D., McCandlish, S., … Amodei, D. (2022). Language models (mostly) know what they know. arXiv. arXiv:2207.05221.
Kambhampati, S. (2024). LLMs can't plan, but can help planning in LLM-modulo frameworks. arXiv. arXiv:2402.01817.
Kambhampati, S. (2024). Can large language models reason and plan? Annals of the New York Academy of Sciences, 1534(1), 15–18. https://doi.org/10.1111/nyas.15125.
Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3128–3137).
Ketokivi, M., & Mantere, S. (2010). Two strategies for inductive reasoning in organizational research. Academy of Management Review, 35(2), 315–333. https://doi.org/10.5465/AMR.2010.48463336
Kim, W. C., & Mauborgne, R. (2005). Blue ocean strategy: How to create uncontested market space and make the competition irrelevant. Harvard Business School Press.
Kirzner, I. M. (1973). *Competition and entrepreneurship.* University of Chicago Press.
Knight, F. H. (1921). *Risk, uncertainty, and profit*. Houghton Mifflin.

Koellinger, P., Minniti, M., & Schade, C. (2007). “I think I can, I think I can”: Overconfidence and entrepreneurial behavior. Journal of Economic Psychology, 28(4), 502–527. https://doi.org/10.1016/j.joep.2006.11.002
Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In Advances in Neural Information Processing Systems (Vol. 35, pp. 22199–22213).
Kuhn, T. S. (1962). The structure of scientific revolutions. University of Chicago Press.
Kurzweil, R. (2022). *The singularity is nearer.* Viking.
Laban, P., et al. (2025). LLMs get lost in multi-turn conversation. arXiv. arXiv:2505.06120.
Lamb, A., He, D., Goyal, A., Ke, G., Ravanelli, M., Bengio, Y., et al. (2021). Transformers with competitive ensembles of independent mechanisms. arXiv. arXiv:2103.00336.
LeCun, Y. (2022, September 9). AI and the limits of language [LinkedIn post]. LinkedIn. https://www.linkedin.com/posts/yann-lecun_ai-and-the-limits-of-language-activity-6967929903409205248--ypi
LeCun, Y. (2022). A path towards autonomous machine intelligence. OpenReview position paper.
Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, e253. https://doi.org/10.1017/S0140525X16001837
Lakoff, G., & Johnson, M. (1980). Metaphors we live by. University of Chicago Press.
Lakoff, G., & Johnson, M. (1999). Philosophy in the flesh: The embodied mind and its challenge to Western thought. Basic Books.
Lampinen, A. K., Dasgupta, I., Chan, S., et al. (2024). Language models, like humans, show content effects on reasoning tasks. PNAS Nexus, 3(7), pgae233. https://doi.org/10.1093/pnasnexus/pgae233
Leibniz, G. W. (1704/1996). New essays on human understanding (P. Remnant & J. Bennett, Trans.). Cambridge University Press. See Book II, Chapter ix, §8.
Lehman, J., Meyerson, E., El-Gaaly, T., Stanley, K. O., & Ziyaee, T. (2025, January 22). Evolution and the Knightian blindspot of machine learning. arXiv. arXiv:2501.13075.
Levinthal, D. A. (1997). Adaptation on rugged landscapes. Management Science, 43(7), 934–950. https://doi.org/10.1287/mnsc.43.7.934
Lewis, M., Feng, S., & Mitchell, M. (2024). Using counterfactual tasks to evaluate the generality of analogical reasoning in LLMs. arXiv. arXiv:2402.08955.
Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022) (pp. 3214–3229). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229
Ling, W., Yogatama, D., Dyer, C., & Blunsom, P. (2017). Program induction by rationale generation: Learning to solve and explain algebraic word problems. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 158–167). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1015
Liu, H., Teng, Z., Ning, R., Ding, Y., Li, X., Liu, X., & Zhang, Y. (2025). GLoRE: Evaluating logical reasoning of large language models (arXiv No. 2310.09107v2) [Preprint]. arXiv.
Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., & Zhang, Y. (2020). LogiQA: A challenge dataset for machine reading comprehension with logical reasoning (arXiv No. 2007.08124) [Preprint]. arXiv.
Liu, T., Xu, W., Huang, W., Zeng, Y., Wang, J., Wang, X., Yang, H., & Li, J. (2025). Logic-of-Thought: Injecting logic into contexts for full reasoning in large language models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers, pp. 10168–10185). Association for Computational Linguistics.
Liu, X., Chen, T., Da, L., et al. (2025b). Uncertainty quantification and confidence calibration in LLMs: A survey. ACM Computing Surveys (in press); arXiv:2503.15850.
Locke, J. (1690/1975). An essay concerning human understanding (P. H. Nidditch, Ed.). Oxford University Press. (Original work published 1690).

LongBench v2 Team. (2025). LongBench v2: A comprehensive long-context benchmark. Project whitepaper.
Loureiro, A. de S., Valverde-Rebaza, J., Noguez, J., Escarcega, D., & Marcacini, R. (2025). Advancing multi-step mathematical reasoning via multi-layered self-reflection with auto-prompting (MAPS). arXiv:2506.23888.
Lovallo, D., Clarke, C., & Camerer, C. (2011). Robust analogizing and the outside view: Two empirical tests of case-based decision making. Strategic Management Journal, 33(5), 496–512.
Makridakis, S., Hogarth, R. M., & Gaba, A. (2010). Why forecasts fail—and what to do instead. MIT Sloan Management Review, 51(2), 83–90.
March, J. G. (1991). Exploration and exploitation in organizational learning. Organization Science, 2(1), 71–87.
March, J. G., & Simon, H. A. (1958). *Organizations*. Wiley.
Marcus, G. (2022). Deep learning is hitting a wall. IEEE Spectrum, August 2022. spectrum.ieee.org/deep-learning-limitations
Marcus, G., & Davis, E. (2020). Rebooting AI: Building artificial intelligence we can trust. Vintage.
Martin, R. L. (2009). The design of business: Why design thinking is the next competitive advantage. Harvard Business Press.
McBride, R., Packard, M. D., & Clark, B. B. (2024). Rogue entrepreneurship. Entrepreneurship Theory and Practice, 48(1), 392–417. doi:10.1177/10422587221135763
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. arXiv:2309.13638v1.
McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 5(4), 115–133.
Meta. (2024). Meta-Llama-3.1-70B-Instruct: Model card [Large language model]. Hugging Face.
Mintzberg, H. (1994). The rise and fall of strategic planning. Free Press.
Mintzberg, H., & Waters, J. A. (1985). Of strategies, deliberate and emergent. Strategic Management Journal, 6(3), 257–272.
Mitchell, M. (2021). Artificial intelligence: A guide for thinking humans. Penguin.
Minsky, M. (1974). A framework for representing knowledge. MIT Artificial Intelligence Laboratory Memo 306.
Molyneux, W. (1688). A letter to the Honourable Robert Boyle. In J. Locke, An essay concerning human understanding(1690/1975, Book II, Chapter IX, Section 8). Oxford University Press.
Mossberg, W. (2007, June 27). The iPhone matches rivals' features, but its main strength may be ease of use. *Wall Street Journal*.
Nasr, M., Rando, J., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Tramèr, F., & Lee, K. (2025). Scalable extraction of training data from aligned, production language models. In Proceedings of the International Conference on Learning Representations (ICLR 2025).
Nelson, R. R., & Winter, S. G. (1982). An evolutionary theory of economic change. Harvard University Press.
Nonaka, I., & Takeuchi, H. (1995). The knowledge-creating company: How Japanese companies create the dynamics of innovation. Oxford University Press.
OpenAI. (2023). GPT-4 technical report. arXiv:2303.08774.
OpenAI. (2024). Learning to reason with large language models. OpenAI Research Blog.
Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback.Advances in Neural Information Processing Systems, 35, 27730–27744.
Pais, A. (1982). Subtle is the Lord: The science and the life of Albert Einstein. Oxford University Press.
Pearl, J. (2000). Causality: Models, reasoning, and inference. Cambridge University Press.

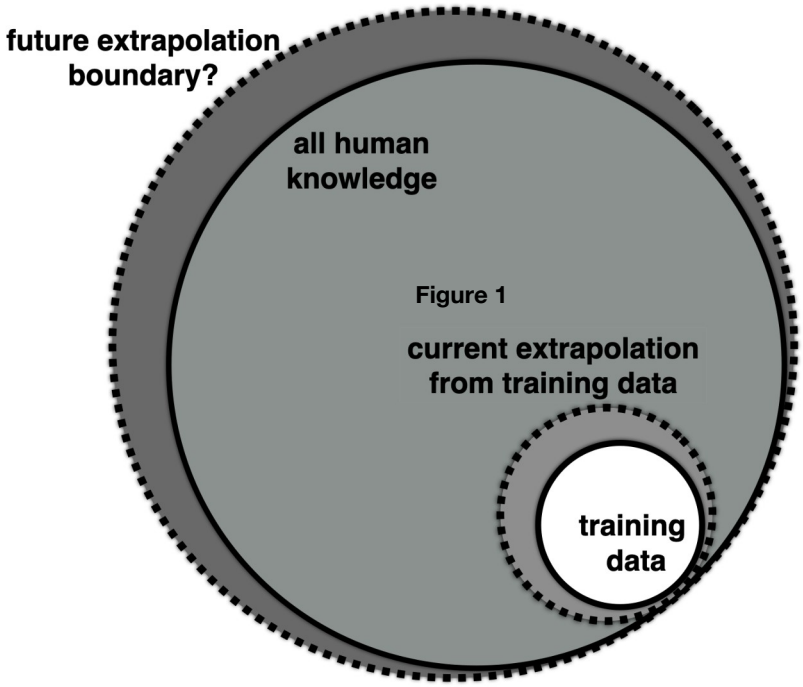
Peirce, C. S. (1955). Philosophical writings of Peirce (J. Buchler, Ed.). Dover Publications.
Peteraf, M. A. (1993). The cornerstones of competitive advantage: A resource-based view. Strategic Management Journal, 14(3), 179–191.
Poddar, A., Sahoo, S., & Ghosh, S. (2025). Understanding syllogistic reasoning in LLMs from formal and natural language perspectives (arXiv No. 2512.12620v3) [Preprint]. arXiv.
Polanyi, M. (1966). The tacit dimension. University of Chicago Press.
Poli, M., Massaroli, S., Nguyen, E. N., Dao, T., Ermon, S., Ré, C., & Bengio, Y. (2023). Hyena hierarchy: Towards larger convolutional language models. arXiv:2302.10866.
Porter, M. E. (1980). Competitive strategy: Techniques for analyzing industries and competitors. Free Press.
Porter, M. E. (1996). What is strategy? Harvard Business Review, 74(6), 61–78.
Porter, M. E. (2008). The five competitive forces that shape strategy. Harvard Business Review, 86(1), 25–40.
Posen, H. E., Keil, T., Kim, S., & Meissner, F. (2018). Renewing research on problemistic search: A review and research agenda. Academy of Management Annals, 12(1), 208–251.
Powell, T. C., Lovallo, D., & Fox, C. R. (2011). Behavioral strategy. Strategic Management Journal, 32(13), 1369–1386.
Pulvermüller, F. (2005). Brain mechanisms linking language and action. Nature Reviews Neuroscience, 6(7), 576–582. doi:10.1038/nrn1706
Rahmandad, H., & Sterman, J. D. (2012). Reporting guidelines for simulation-based research in the social sciences. System Dynamics Review, 28(4), 396–411.
Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. Academy of Management Review, 46(1), 192–210.
Raza, M., & Milic-Frayling, N. (2025). Instantiation-based formalization of logical reasoning tasks using language models and logical solvers. In Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25) (pp. 4633–4641). International Joint Conferences on Artificial Intelligence Organization.
Ren, R., Xie, S. M., et al. (2024). Toward understanding how transformers learn in context. In Advances in Neural Information Processing Systems (NeurIPS 2024).
Repenning, N. P. (2002). A simulation-based approach to understanding the dynamics of innovation implementation. Organization Science, 13(2), 109–127.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536. https://doi.org/10.1038/323533a0
Rumelt, R. P. (1991). How much does industry matter? Strategic Management Journal, 12(3), 167–185.
Sadr, N. G., Madhusudan, S., Sajjad, H., & Emami, A. (2025). Which words matter most in zero-shot prompts? (arXiv No. 2502.03418v3) [Preprint]. arXiv.
Saparov, A., & He, H. (2022). Language models are greedy reasoners: A systematic formal analysis of chain-of-thought (arXiv No. 2210.01240) [Preprint]. arXiv.
Schendel, D. E., & Hofer, C. W. (1979). Strategic management: A new view of business policy and planning. Little, Brown.
Schoemaker, P. J. H. (1995). Scenario planning: A tool for strategic thinking. Sloan Management Review, 36(2), 25–40.
Schumpeter, J. A. (1934). The theory of economic development. Harvard University Press.
Seals, S., et al. (2024). Evaluating the deductive competence of large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024).
Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–424.
Searle, J. R. (1983). Intentionality: An essay in the philosophy of mind. Cambridge University Press.
Searle, J. R. (1990). Is the brain’s mind a computer program? Scientific American, 262(1), 26–31.
Senge, P. M. (1990). The fifth discipline: The art and practice of the learning organization. Doubleday/Currency.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL).
Sergeeva, A., Bhardwaj, A., & Dimov, D. (2021). In the heat of the game: Analogical abduction in a pragmatist account of entrepreneurial reasoning. Journal of Business Venturing, 36(6), 106158. doi:10.1016/j.jbusvent.2021.106158
Shepherd, D. A., & Majchrzak, A. (2022). Machines augmenting entrepreneurs: Opportunities (and threats) at the nexus of artificial intelligence and entrepreneurship. Journal of Business Venturing, 37(5), 106227.
Shoemaker, P. J. H. (1996). Scenario planning: A tool for strategic thinking. Sloan Management Review, 36(2), 25–40.
Simon, H. A. (1957). Models of man: Social and rational. John Wiley & Sons.
Sinha, K., Sodhani, S., Dong, J., Pineau, J., & Hamilton, W. L. (2019). CLUTRR: A diagnostic benchmark for inductive reasoning from text (arXiv:1908.06177) [Preprint]. arXiv.
Solow, R. M. (1987). We'd better watch out. New York Times Book Review, July 12.
Sprague, Z., et al. (2023). MuSR: Testing the limits of chain-of-thought (murder-mystery abduction) (arXiv:2310.16049) [Preprint]. arXiv.
Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models (arXiv:2206.04615v3) [Preprint]. arXiv.
Sterman, J. D. (1989). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. Management Science, 35(3), 321–339.
Sterman, J. D. (2000). Business dynamics: Systems thinking and modeling for a complex world. Irwin/McGraw-Hill.
Su, E., Vellore, A., Chang, A., Mura, R., Nelson, B., Kassianik, P., & Karbasi, A. (2024). Extracting memorized training data via decomposition (arXiv:2409.12367) [Preprint]. arXiv.
Sun, H., Min, Y., Chen, Z., Zhao, W. X., Liu, Z., Wang, Z., Fang, L., & Wen, J.-R. (2025). Challenging the boundaries of reasoning: An Olympiad-level math benchmark for large language models (OlymMATH) (arXiv:2503.21380) [Preprint]. arXiv.
Surowiecki, J. (2008). *The wisdom of crowds.* Anchor Books.
Sutton, R. (2019, March 13). The bitter lesson. Incomplete Ideas. incompleteideas.net/IncIdeas/BitterLesson.html
Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., & Wei, J. (2022). Challenging BIG-Bench tasks and whether chain-of-thought can solve them (arXiv:2210.09261) [Preprint]. arXiv.
Tafjord, O., Dalvi, B., & Clark, P. (2021). ProofWriter: Generating implications, proofs, and abductive statements over natural language. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 3621–3634). Association for Computational Linguistics.
Tang, Z., et al. (2024). Human-like cognitive patterns as emergent phenomena in LLMs (arXiv:2412.15501) [Preprint]. arXiv.
Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. Strategic Management Journal, 18(7), 509–533.
Thompson, C. (2012). Smarter than you think: How technology is changing our minds for the better. Penguin Press.
Togelius, J., & Yannakakis, G. N. (2024). Choose your weapon: Survival strategies for depressed AI academics. Proceedings of the IEEE, 112(1), 4–11. doi:10.1109/JPROC.2024.3364137
Uehara, S. (2025, September 4). 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., Marquez, M., Sreedharan, S., & Kambhampati, S. (2023). On the planning abilities of large language models: A critical investigation. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in Neural Information Processing Systems (Vol. 36). Neural Information Processing Systems Foundation.
Valmeekam, K., Stechly, K., & Kambhampati, S. (2024). LLMs still can't plan; can LRMs? A preliminary evaluation of OpenAI's o1 on PlanBench (arXiv:2409.13373) [Preprint]. arXiv. doi:10.48550/arXiv.2409.13373
Varela, F. J., Thompson, E., & Rosch, E. (1991). The embodied mind: Cognitive science and human experience. MIT Press.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
Vogelstein, F. (2013). Dogfight: How Apple and Google went to war and started a revolution. Farrar, Straus and Giroux.
von Oswald, J., Niklasson, E., Randazzo, E., et al. (2022). Transformers learn in-context by gradient descent. In Proceedings of the International Conference on Learning Representations (ICLR 2023).
Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., & Zhou, D. (2022). Self-consistency improves chain-of-thought reasoning. In Proceedings of the International Conference on Learning Representations (ICLR 2023).
Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), New horizons in psychology (pp. 135–151). Penguin Books.
Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. Nature Human Behaviour, 7, 1526–1541.
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models (arXiv:2201.11903) [Preprint]. arXiv.
Weick, K. E. (1989). Theory construction as disciplined imagination. Academy of Management Review, 14(4), 516–531.
Weick, K. E. (1995). Sensemaking in organizations. Sage Publications.
Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., & Weston, J. (2019). Neural text generation with unlikelihood training. In Proceedings of the International Conference on Learning Representations (ICLR).
Wenger, E., & Kenett, Y. N. (2025). We're different, we're the same: Creative homogeneity across LLMs (arXiv:2501.19361) [Preprint]. arXiv.
Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic Management Journal*, 5(2), 171–180.
Whittaker, E. T. (1951). A history of the theories of aether and electricity (Vol. 2). Thomas Nelson.
Wooldridge, M. (2018). *A brief history of artificial intelligence: What it is, where we are, and where we are going.* Flatiron Books.
Wooldridge, M. (2020). A brief history of artificial intelligence. Flatiron Books.
Wu, D., Yang, J., & Wang, K. (2024). Exploring the reversal curse and other deductive challenges in large language models. Patterns, 5(10), 100945. https://doi.org/10.1016/j.patter.2024.100945.
Xie, S. M., Raghunathan, A., Liang, P., & Ma, T. (2022). In-context learning as implicit Bayesian inference. In Proceedings of the International Conference on Learning Representations (ICLR 2022).
Yang, Z., et al. (2024). MOOSE: Multi-step open-domain scientific exploration. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024).
Yang, Z., et al. (2024). MOOSE-Chem: Large language models for rediscovering unseen chemistry scientific hypotheses. OpenReview.
Yu, W., Jiang, Z., Dong, Y., & Feng, J. (2020). ReClor: A reading comprehension dataset requiring logical reasoning. In Proceedings of the International Conference on Learning Representations (ICLR).
Zečević, M., Willig, M., Dhami, D. S., & Kersting, K. (2023). Causal parrots: Large language models may talk causality but are not causal. Transactions on Machine Learning Research, 2023(8).

Zhang, W., Xu, W., Liu, Y., Li, Y., & Sun, M. (2024). Pretraining data detection for large language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024) (pp. 3567–3580). Association for Computational Linguistics.
Zhong, W., Wang, S., Tang, D., Xu, Z., Guo, D., Chen, Y., Wang, J., Yin, J., Zhou, M., & Duan, N. (2022). Analytical reasoning of text. In Findings of the Association for Computational Linguistics: NAACL 2022 (pp. 2306–2319). Association for Computational Linguistics.
Zhou, J., Ghaddar, A., Zhang, G., Ma, L., Hu, Y., Pal, S., Coates, M., Wang, B., Zhang, Y., & Hao, J. (2024). Enhancing logical reasoning in large language models through graph-based synthetic data (arXiv No. 2409.12437v2) [Preprint]. arXiv.
Zhou, J., Sinha, M., Dong, J., Li, Y., & Hu, Y. (2024). Enhancing Logical Reasoning in Large Language Models via Extract-Then-Answer Prompting. arXiv preprint arXiv:2409.12437. https://doi.org/10.48550/arXiv.2409.12437.
Zhou, Y., Ye, J., Ling, Z., Han, Y., Huang, Y., Zhuang, H., Liang, Z., Guo, K., Guo, T., Wang, X., & Zhang, X. (2025). Dissecting logical reasoning in large language models: A fine-grained evaluation and supervision study (arXiv:2506.04810) [Preprint]. arXiv.

Figures



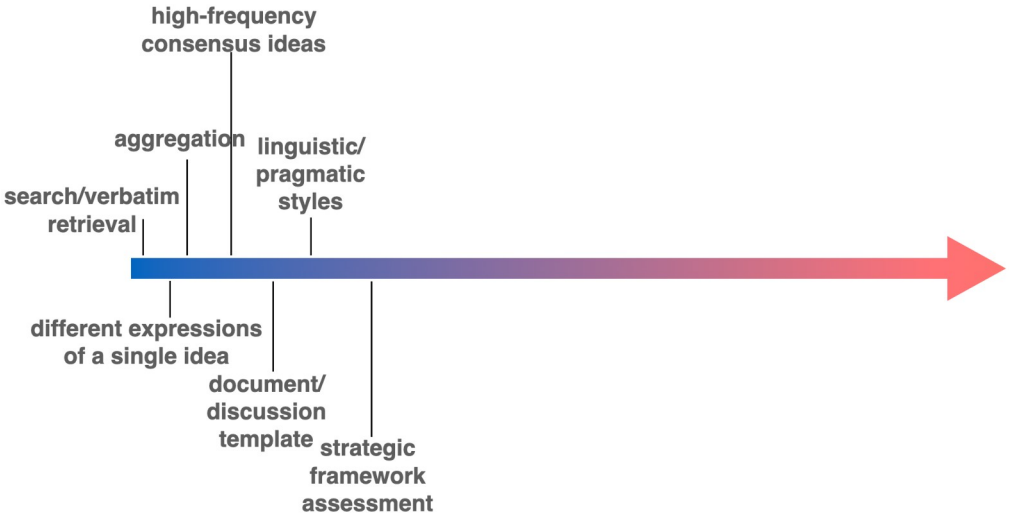

high-frequency
consensus ideas
aggregation
linguistic/
pragmatic
styles
search/verbatim
retrieval
different expressions
of a single idea
document/
discussion
template
strategic
framework
assessment

Figure 2

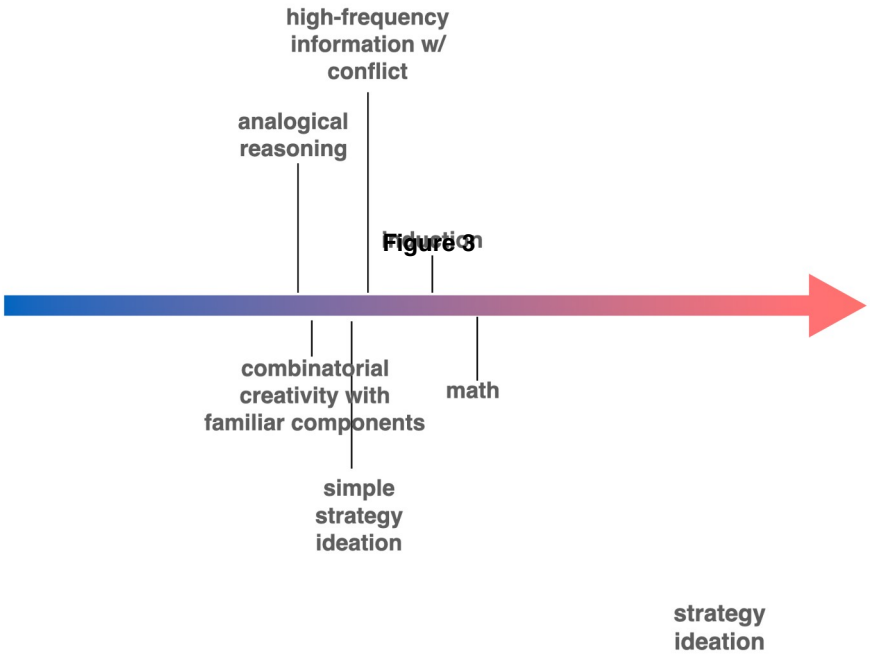


Figure 3

Figure 4

Figure 5

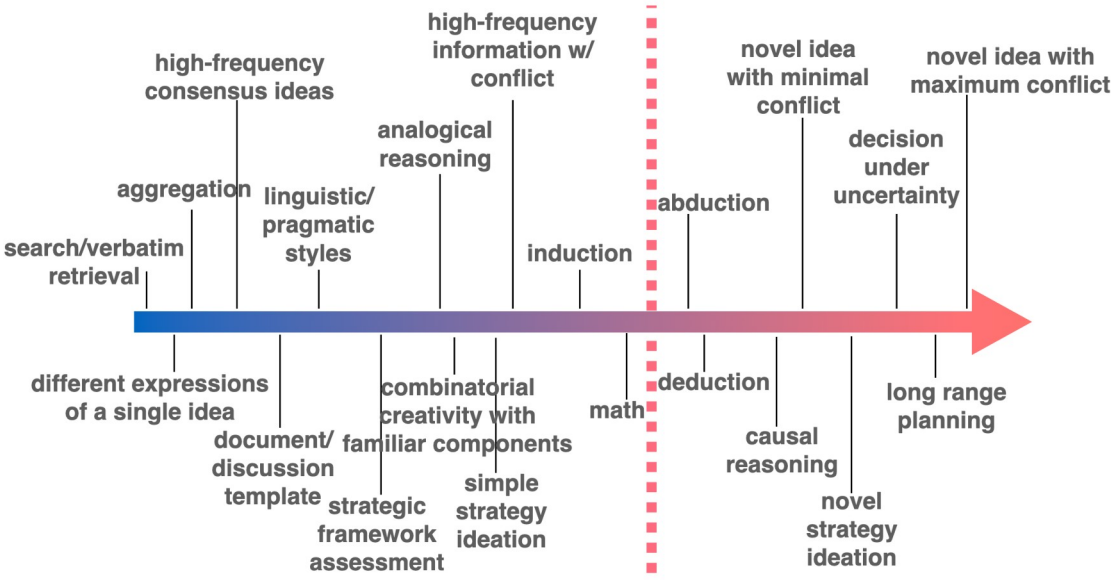

high-frequency consensus ideas
high-frequency information w/ conflict
novel idea with minimal conflict
novel idea with maximum conflict
analogical reasoning
aggregation
linguistic/ pragmatic styles
decision under uncertainty
abduction
search/verbatim retrieval
induction
different expressions of a single idea
document/ discussion template
combinatorial creativity with familiar components
math
deduction
long range planning
causal reasoning
strategic framework assessment
simple strategy ideation
novel strategy ideation

Figure 6

Appendix A— Key Advances

** indicates the most important advances

Year	Technology	Advance	Inventor(s)
1986	Backpropagation	Enabled multi-layer neural network training via gradient descent.	Rumelhart, Hinton, & Williams
2014	Adam Optimizer	Improved optimization by combining momentum and adaptive learning rates.	Kingma & Ba
2017	**Transformer Architecture	Introduced self-attention and parallelization for scalable language modeling.	Vaswani et al.
2017	**Self-Attention Mechanism	Enabled contextual understanding by allowing each token to attend to all other tokens in a sequence.	Vaswani et al.
2018	Unsupervised Pretraining (ULMFiT)	Demonstrated transfer learning via language model fine-tuning.	Howard & Ruder
2018	GPT-1 (OpenAI)	Introduced generative pretraining for NLP tasks.	Radford et al.
2018	BERT (Google)	Introduced bidirectional context via masked language modeling.	Devlin et al.
2020	**Scaling Laws	Established empirical laws showing predictable improvements with scale.	Kaplan et al.
2020	GPT-3 (OpenAI)	Scaled transformer-based language modeling to 175 billion parameters, demonstrating few-shot learning.	Brown et al.
2022	*Human Instruction Tuning (InstructGPT)	Aligned LLM outputs with human preferences using supervised fine-tuning and reinforcement learning from human feedback (RLHF).	Ouyang et al.

Appendix B

A few recommended sources for introductory learning material related to AI

AI Tutorials—
3Blue1Brown.com has excellent tutorials on introductory neural networks and related math:
https://www.3blue1brown.com/topics/neural-networks

The gold standard for deep learning courses are Andrew Ng's Coursera courses (and, as a side note Coursera's beginnings are analyzed in Volmer & Eisenhardt (2025) under the pseudonym, 'Maverick'):
https://www.coursera.org/specializations/deep-learning

Good Technical Programming Books on Machine Learning—
Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

Good Accessible Overview of AI—
Mitchell, M. (2021). *Artificial intelligence: A guide for thinking humans*. Penguin Books.

7 Argumentation Maps that cover every (!) major philosophical debate around A.I.
Horn, Robert, Jeff Yoshimi, Mark Deering, and Russ McBride. 1998. Can Computers Think: The History and Status of the Debate. A series of 7 maps and a handbook. MacroVu Inc. BainBridge, WA.
Links to PDFs of the maps:
1— http://www.bobhorn.us/assets/cct-map-1-whatisthis-1998_reduced2.pdf
2— http://www.bobhorn.us/assets/cct-map-2-whatisthis-1998_reduced.pdf
3— https://www.bobhorn.us/assets/cct-map-3-whatisthis-1998_reduced.pdf
4— http://www.bobhorn.us/assets/cct-map-4-whatisthis-1998_reduced.pdf
5— https://www.bobhorn.us/assets/cct-map-5-whatisthis-1998_reduced.pdf
6— https://www.bobhorn.us/assets/cct-map-6-whatisthis-1998_reduced.pdf
7— http://www.bobhorn.us/assets/cct-map-7-whatisthis-1998_reduced.pdf

Appendix C

Deductive Reasoning Benchmark Performance

Benchmark	Year	Task Type	Metric	Human %	Best 2025 LLM %	Best Model	Evaluated Regime	Delta % (Human–LLM)	Citations
ReClor	2020	Logical reasoning (MC)	Accuracy	100.0*	94	QwQ-32B	GLoRE v2 reported result (audit line)	6	(Yu et al., 2020; Liu et al., 2025b)
LogiQA	2020	Logical RC (MC)	Accuracy	95.0*	62	GPT-4-0613	CoT-SC(5), zero-shot; mixed Zh+En test selection	33	(Liu et al., 2020; Liu et al., 2025a)
AR-LSAT	2022	Analytical reasoning	Accuracy	91.0*	71	GPT-4 + solver (SSV)	Semantic self-verification + solver (SSV)	20	(Raza & Milic-Frayling, 2025; Zhong et al., 2022; Liu et al., 2023)
ProofWriter	2021	Symbolic logic	Accuracy	93.0*	98	GPT-4 + solver (SSV)	Semantic self-verification + solver (SSV)	-5	(Raza & Milic-Frayling, 2025; Tafjord et al., 2021; Liu et al., 2023)
FOLIO	2022/2023	First-order logic	Accuracy	100.0**	81	GPT-4 + solver (SSV)	Semantic self-verification + solver (SSV)	19	(Han et al., 2022/2023; Raza & Milic-Frayling, 2025)
EntailmentBank	2021	Multi-step deduction	Accuracy (mean over hops 1–6)	94.0*	—	—	No single citable “mean accuracy over hops 1–6” under stated regime in cited sources; requires separate audit	—	(Dalvi et al., 2021; Cheng et al., 2024)
BIG-Bench Logical Deduction (hard)	2022	Multi-step logic	Accuracy	100.0*	90	GPT-4 + solver (SSV)	Semantic self-verification + solver (SSV)	10	(Raza & Milic-Frayling, 2025)
PrOntoQA	2021–2022	Ontological reasoning	Accuracy	100.0*	100	GPT-4 + solver (SSV)	Semantic self-verification + solver (SSV)	0	(Raza & Milic-Frayling, 2025)
CLUTRR (2-hop)	2019	Relational deduction	Accuracy	100.0*	98	GPT-4o	Few-shot + ETA-P prompting (reported)	2	(Sinha et al., 2019)
CLUTRR (6-hop)	2019	Relational deduction	Accuracy	100.0*	88	GPT-4o	Few-shot + ETA-P prompting (reported)	12	(Sinha et al., 2019)
DROP	2019	Discrete reasoning	F1	96	85	Llama 3.1 405B Instruct	Model-card reported DROP(F1)	12	(Dua et al., 2019; Meta, 2024)

RuleTaker	2020	Rule-based deduction	Accuracy	95.0*	71	GPT-4-0613	LoT table setup + CoT-SC(5) (depth-5 CWA subset)	24	(Clark et al., 2020; Liu et al., 2025a)
MathQA	2019	Logic + math word problems	Accuracy	100.0*	54	Seq2prog+cat (non-LLM baseline)	Supervised baseline reported in dataset work (not an LLM)	46	(Amini et al., 2019)
AQuA-RAT	2017	Algebraic reasoning (MC)	Accuracy	100.0*	68	GPT-4o mini	CoT prompt condition reported in study	32	(Ling et al., 2017; Sadr et al., 2025)
Syllogistic reasoning	2025	Formal logic	Accuracy	100.0*	100	Gemini 2.5 Flash	Mean of congruent/incongruent accuracies in reported setup	0	(Poddar et al., 2025)

Human and model performance on logical/deductive reasoning benchmarks: Human values are reported baselines when available; otherwise they reflect an explicitly marked ceiling-style reference point (e.g., logical ceiling or expert ceiling) rather than a measured population mean. “Best 2025 LLM %” reports the highest citable score verified for a specified model under the stated evaluation regime (prompting-only vs solver/tool-augmented, subset/depth/hop constraints, and metric definition). Because benchmarks differ in splits, metrics (accuracy vs F1), and allowed inference procedures, the human–model deltas should be interpreted as within-row gaps under the listed protocol rather than as a single unified cross-benchmark leaderboard.

Δ is computed as Human − LLM in percentage points using the displayed numeric values (one-decimal precision).
Metrics are not directly comparable across rows (e.g., accuracy vs F1 vs exact match vs proof-F1/tree correctness); comparisons should be interpreted only within the row.

Solver/tool-augmented rows (e.g., SSV) reflect system performance (LLM + external verifier/solver) and should not be interpreted as “LLM-only” prompting performance.

“—” indicates that a single numeric value is not citable under the stated metric/regime from the listed sources and would require a separate protocol-locked audit.

* = “Human %” is an ideal/ceiling comparator (logical ceiling or expert ceiling) used when a measured human baseline for the same split/protocol is not reported.
*
** = FOLIO is formally annotated/logic-grounded, but “100% human” is not a standard reported baseline for the evaluated subset—use only if explicitly labeled “logical ceiling.”

Disambiguation note: “Liu et al., 2025a” refers to the Logic-of-Thought paper; “Liu et al., 2025b” refers to the GLoRE v2 audit paper.