From Statistical Pattern Matching to Deliberative Inference: A Critical Review of Reasoning and Faithfulness in Large Language Models

Lohith Srikanth Pentapalli, Kaaustaaub Shankar, Atharv Shete

April 2026

Abstract

The rapid advancement of large language models (LLMs), particularly those based on transformer architectures, has reignited fundamental questions about the nature of machine intelligence and reasoning. Recent developments such as Chain-of-Thought (CoT) prompting, self-consistency, and structured reasoning frameworks (e.g., Tree-of-Thought and graph-based approaches) have demonstrated substantial improvements in performance on complex, multi-step tasks. These advances have led to growing claims that modern LLMs exhibit forms of deliberative reasoning approaching human-like cognition.

This review critically examines these claims by analyzing the distinction between reasoning as observed in model outputs and reasoning as an internal, causal process. While LLMs can generate coherent and interpretable reasoning traces, emerging evidence suggests that such outputs may be unfaithful to the underlying decision-making mechanisms, raising concerns about interpretability and misplaced trust. Furthermore, recent methods that rely on sampling, search, and external evaluation complicate the attribution of reasoning to the model itself, instead framing performance gains as a product of inference-time strategies.

By synthesizing foundational and contemporary research, this paper argues that although LLMs exhibit increasingly sophisticated reasoning-like behaviors, the extent to which these reflect genuine understanding remains unresolved. This distinction has significant implications for the deployment of LLMs in high-stakes domains, where reliability, transparency, and accountability are critical.

Keywords: Large Language Models, Reasoning, Chain-of-Thought, Tree-of-Thought, Interpretability, Explainability, Artificial Intelligence, Deliberative Reasoning, Inference-Time Computation, Trustworthiness

1 Introduction

Understanding intelligence has been a critical objective in science for quite some time. In 1950, Alan Turing posed one of the field's most enduring questions in

his seminal paper, "Computing Machinery and Intelligence," proposing that machine intelligence could be evaluated through what he called the imitation game, later known as the Turing Test [1]. With the emergence of computers in the mid-20th century, recreating intelligence has become an increasingly tractable problem. The field of Artificial Intelligence (AI) was formally established at the Dartmouth Conference in 1956, where researchers set out to explore how machines could use language, form abstractions, and solve problems traditionally reserved for humans [2]. The earliest use of AI relied on symbolic logic, which helped represent knowledge through rules and logical principles [3][4].

For decades, much of the progress in the field was incremental and focused around certain tasks like games [5][6] and image classification [7]. This changed with the introduction of pre-trained transformers [8], along with the advent of GPT, a class of generative pre-trained transformer models developed by OpenAI. Beginning with GPT-2 [9] in 2019 and escalating with GPT-3 [10] in 2020, these models demonstrated unprecedented scaling properties, where larger models exhibited emergent capabilities not present in smaller versions [10]. The release of ChatGPT on November 30, 2022, marked a significant inflection point by bringing these capabilities to the public through a conversational interface [11]. Unlike its predecessors, ChatGPT is a large language model (LLM) capable of seemingly intelligent conversation across diverse domains, achieving rapid adoption with over one million users within five days of its public release [12]. Built on the GPT-3.5 architecture (later upgraded to GPT-4), ChatGPT differs fundamentally from earlier language models by being specifically optimized for natural dialogue and instruction-following through reinforcement learning from human feedback [13].

This breakthrough has reignited long-standing questions about machine intelligence and reasoning. Some researchers have argued that systems like GPT-4 demonstrate capabilities spanning mathematics, coding, medicine, law, and psychology, suggesting an early form of artificial general intelligence [14]. Indeed, recent work indicates that ChatGPT exhibits improved reasoning abilities, including proficiency in chain-of-thought reasoning that mirrors human deliberative thinking [15]. These capabilities have led to claims that LLMs achieve genuine reasoning and problem-solving abilities. However, this optimistic view deserves scrutiny. Despite strong benchmark performance, key questions remain: do LLMs truly reason or just mimic patterns? Does scaling produce real understanding or better imitation? As their use expands into high-stakes fields like healthcare and law, issues of interpretability, hallucinations, and reliability become critical.

This review systematically examines these questions through a critical lens. We trace the evolution from early AI systems to modern LLMs, evaluate competing claims about reasoning abilities, assess recent advances including reinforcement learning approaches, and ultimately address whether current systems should be trusted in safety-critical applications. Our analysis suggests that while LLMs have achieved remarkable capabilities, the nature of what they do and what they truly understand remains ambiguous.

2 From Patterns to Reasoning

One of the main reasons large language models began to be discussed as possible reasoning systems was the success of Chain of Thought (CoT) prompting. Wei et al. showed that when sufficiently large models are prompted with intermediate reasoning steps, performance improves substantially on arithmetic, common-sense, and symbolic-reasoning tasks [15]. This was important because it suggested that models could benefit from breaking complex problems into smaller steps rather than directly predicting a final answer. In many cases, the resulting outputs looked more structured, more interpretable, and more similar to the kind of step-by-step reasoning humans often use when solving difficult problems. This led many researchers to view CoT as an early sign that large models might be capable of something closer to deliberative reasoning, specifically informal deductive reasoning, rather than simple surface-level pattern completion.

However, the theoretical interpretation of these results is less straightforward than it first appears. CoT clearly improves performance and mimics human reasoning, but that alone does not establish that a model is actually reasoning in a human-like or causal sense [15]. The model is still generating text auto-regressively, one token at a time, and the fact that this text contains intermediate steps does not necessarily mean those steps faithfully reflect the process that produced the answer. In other words, CoT may be useful without being fully explanatory. This distinction becomes especially important in review of the broader literature, because much of the field's early optimism came from observing reasoning-like behavior at the level of output, while the internal process remained largely opaque.

Several follow-up works expanded on the basic CoT idea and showed that reasoning performance could be further improved through better inference or training strategies. Kojima et al. augmented prompts with a simple "Think step by step" phrase, which also improved performance in arithmetic and symbolic reasoning tasks [16]. This led to the theory that LLMs possess significant untapped zero-shot knowledge and cognitive abilities that can be extracted through simple instructions rather than extensive human-engineered examples. Fu et al. proposed complexity-based prompting, which creates longer chains of thoughts to improve performance [17]. Wang et al. proposed self-consistency, which replaces greedy decoding with a procedure that samples multiple reasoning paths and then selects the most consistent final answer [18]. This produced strong gains across several reasoning benchmarks and showed that there may be many valid reasoning paths leading to the same correct answer. Around the same time, Zelikman et al. introduced STaR, a method that allows a model to bootstrap its own rationales by generating reasoning traces, filtering for successful ones, and then fine-tuning on them [19]. Together, these papers broadened the picture. CoT was no longer just a prompting trick. It became part of a larger family of methods aimed at improving multi-step reasoning through sampling, selection, and rationale-based training. At the same time, these methods also complicated the interpretations of reasoning itself. If performance improves be-

cause many candidate chains are sampled and the best answer is recovered or because models are trained to reproduce successful rationales, then it becomes harder to argue that any single visible chain directly captures the model's actual decision process.

This concern is central to more recent critiques of LLM reasoning. Turpin et al. argue that models do not always say what they think, showing that CoT explanations can be systematically unfaithful to the features that actually influenced a prediction [20]. In their experiments, models were affected by biasing features in the prompt, yet their generated explanations often failed to mention those features and instead presented plausible justifications for the final answer. Barez et al. push this concern further by arguing that CoT is not explainability in any strong sense and that verbalized reasoning traces are neither necessary nor sufficient for trustworthy interpretability [21]. These critiques do not mean that CoT is useless. Rather, they suggest that presence of a coherent reasoning trace should not be taken as direct evidence that the model is engaging in transparent, faithful or human-like deliberation. This is especially relevant as such systems are increasingly discussed for use in higher stakes domains, where a convincing explanation may create misplaced trust.

Taken together, the literature presents a more mixed picture than early enthusiasm sometimes suggested. On the one hand, CoT and its extensions clearly improve performance on multi-step tasks and have reshaped how researchers think about inference-time reasoning in LLMs. On the other hand, the relationship between visible reasoning traces and genuine deliberative reasoning remains uncertain. For this reason, later work began to move beyond single linear chains and toward more structured forms of search and evaluation. For example, Tree of Thoughts explicitly frames problem solving as exploration over multiple possible reasoning paths, allowing lookahead, self-evaluation, and backtracking rather than relying on single left-to-right chain [22]. The emergence of such methods reflects a broader shift in the field from asking whether a model can produce a chain of reasoning to asking whether it can actually explore, compare, and revise competing lines of thought in a more deliberate manner.

3 Inference-Time Reasoning Methods

A natural response to the limitations of single-chain reasoning was to move toward more structured search procedures. Tree of Thought (ToT) is a key example of this shift, allowing models to generate multiple intermediate "thoughts", evaluate them, and selectively continue or backtrack instead of following one fixed reasoning path from start to finish [22]. This reframes reasoning as exploration over a space of possible partial solutions, making it better suited for tasks that require lookahead or recovery from early mistakes. Related work extended this idea across languages, suggesting that structured thought exploration can also improve multilingual reasoning rather than remaining tied to English-centric prompting methods [23].

Recent survey work suggests that inference-time reasoning methods are better understood not as one unified move toward deliberation but as a progression from structured search to guided search, to self-improving reasoning pipelines [24]. At a broad level, reasoning oriented methods can be grouped into structure search, reward modeling, self-improvement, macro action, and reinforcement fine-tuning, though the first 3 matter most for the present discussion. Structure-search methods explicity branch over candidate trajectories instead of relying on a single linear rationale. This includes ToT and later search-based variants such as RAP [25], Forest-of-Thought [26] , and TreeBoN [27] which use test-time computation to explore, compare and revise promising paths. A second line of work focuses on scoring intermediate reasoning steps rather than only final answers using verifiers or reward signals to guide refinement as in DIVERSE [28], ORPS [29], and Step-DPO [30]. A third line reuses sampled trajectories to improve future reasoning behavior as in STaR and later MCTS-based self-improvement approaches. Seen this way the post CoT literature is not a single path toward deliberation but a set of distinct mechanisms for expanding, evaluating and reusing reasoning traces.

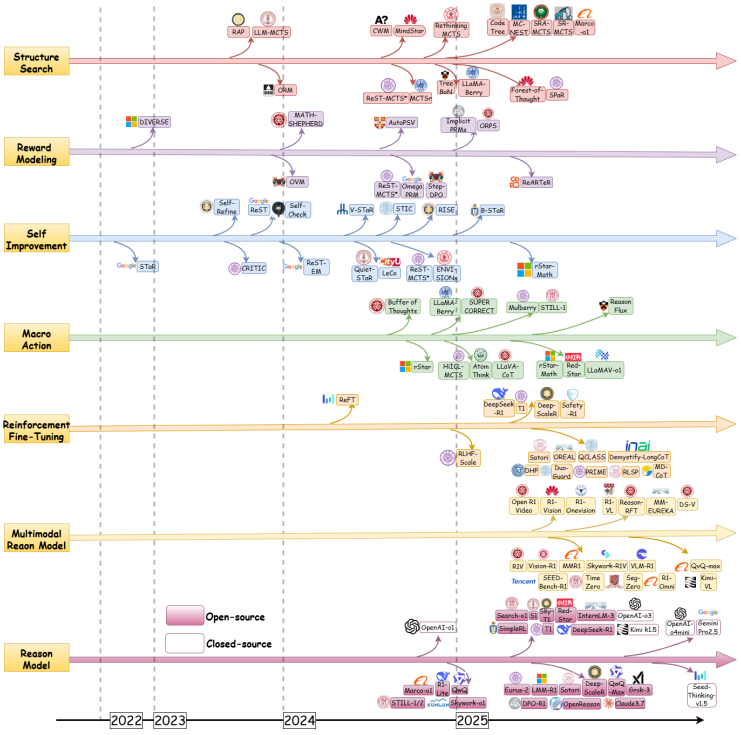

Figure 1: Taxonomy of reasoning-oriented LLM methods over time. Adapted from [24].

Summary of tree search-based methods. **Tag:** N=Natural language, C=Code, E=Math Expression, or A=Action.

Method	Node		Evaluate	Rollout	Tasks	LLM Models
	Format	Partial				
RAP [14]	N	✓	LLM Self-correction	LLM-based Prediction	Planning, Reasoning	LLaMA-33B
ORM [126]	N	✓	Value/Reward Function	N/A	Multiple Tasks	LLaMA-2-7B, GPT-2-small
Forest-of-Thought [127]	N	✓	LLM Self-correction	Self-refinement & Iterative Improvement	Planning, Reasoning	LLaMA-3, Mistral7B, GLM-4-9B
CodeTree [128]	N C	✓	Execution Accuracy + LLM Self-correction	Code Execution	Code Generation	GPT-4
TreeBoN [129]	N	✓	Value/Reward Function	Speculative & Dynamic Strategies	Planning, Reasoning	LLaMA-3-8B
CWM [130]	C	✓	Compare with Golden Data	Code Execution	Alignment Task	N/A
LLM-MCTS [131]	N	✗	LLM Self-correction + Policy	LLM-Based Prediction	Household Environments	GPT-2, GPT-3.5
RethinkMCTS [132]	C	✗	Execution Accuracy	Code Execution	Code Generation	GPT-3.5-turbo, GPT-4o-mini
MCTSr [133]	N	✗	LLM Self-correction	Self-Refinement & Iterative Improvement	Mathematical Reasoning	LLaMA-3-8B
MC-NEST [134]	N	✗	LLM Self-correction	Reasoning Path Generation	Mathematical Reasoning	GPT-4o, Phi-3-mini
SRA-MCTS [135]	N	✓	Execution Accuracy	Reasoning Path Generation	Code Generation	LLaMA-3-70B-Instruct
SPaR [136]	N	✓	LLM Self-correction	Self-Refinement & Iterative Improvement	Instruction Following	LLaMA-3-8B
MindStar [137]	N	✓	Value/Reward Function	Reasoning Path Generation	Mathematical Reasoning	LLaMA-2-13B, Mistral-7B
SR-MCTS [138]		✓	Compare with Golden Data	Math Expression Generation	Financial Fraud Detection	GPT
LLaMA-Berry [139]	N	✗	Compare with Other Solutions	Math Expression Generation	Mathematical Reasoning	LLaMA-3.1-8B
Macro-o1 [112]	N A	✗	LLM's Output Probabilities	Reasoning Path Generation	Multiple Tasks	Qwen2-7B-Instruct
ReST-MCTS* [140]	N	✓	Probability to Correct Answer	Reasoning Path Generation	Mathematical Reasoning	LLaMA-3, Mistral-7B, SciGLM-6B
CoMCTS [141]	N	✓	Compare with Other Solutions	Reasoning Path Generation	Multiple Tasks	GPT-4o, Qwen2-VL-7B, LLaMA-3.2-11B
C-MCTS [142]	N	✓	Compare with Golden Data	Math Expression Generation	Mathematical Reasoning	Qwen2.5
rStar-Math [143]		✓	Compare with Other Solutions	Math Expression Generation	Mathematical Reasoning	Phi3-mini-Instruct, Qwen2.5-Math-1.5B, Qwen2.5-Math-7B
AStar [144]	N	✓	Compare with Other Solutions	Reasoning Path Generation	Multiple Tasks	Qwen2.5-7B, Qwen2-VL-2B, Qwen2-VL-7B
DeepSolution [145]	N A	✗	Compare with Golden Data	Reasoning Path Generation	Multiple Tasks	Qwen2.5-7B-Instruct
VisuoThink [146]	N	✓	Compare with Golden Data	Reasoning Path Generation	Multiple Tasks	GPT-4o, Qwen2-VL-72B-Instruct, Claude-3.5-sonnet
TongGeometry [147]		✓	Compare with Golden Data	Math Expression Generation	Mathematical Reasoning	GPT-4
PPO-MCTS [148]	N	✓	Compare with Other Solutions	Reasoning Path Generation	Alignment Task	LLaMA-7B

Figure 2: Summary of tree search-based methods. Adapted from [24].

Other approaches pushed reasoning even further by combining it with planning, acting, and richer thought structures. ReAct interleaves reasoning traces with actions, allowing a model to gather external information while updating its intermediate decisions [31]. Similarly, Reasoning via Planning and Language Agent Tree Search use explicit search procedures such as Monte Carlo Tree Search to explore alternatives, simulate outcomes, and revise failed paths [25, 32]. At the same time, graph-based frameworks such as Graph of Thoughts argue that even tree structures can be too restrictive for elaborate tasks and instead model reasoning as a more flexible graph of intermediate ideas that can branch, merge, and refine [33]. Together, these methods show that recent work has increasingly treated reasoning not as a single explanation chain, but as a structured process of search, evaluation, and adaptation.

Beyond prompting a single rationale, other work also explored ways to make inference time reasoning more reliable while still keeping it largely linear. Self consistency is one important example. Instead of committing to one reasoning path under greedy decoding, it samples multiple chains and selects the answer that appears most often across them [18]. This does not change the basic form of CoT reasoning but it does make the final prediction less dependent on one possibly fragile intermediate trace. In practice, this often improves performance on arithmetic and symbolic tasks, where different valid reasoning paths may still converge on the same correct answer.

Other work focused less on sampling multiple chains and more on changing how a problem is broken down. Least-to-most prompting addresses cases where standard CoT struggles to generalize from relatively simple examples to more difficult target problems [34]. Rather than asking the model to solve the full task in one continuous explanation, it first decomposes the problem into simple subproblems and then solves them step-by-step. This keeps the reasoning process sequential but makes it more structured by introducing an explicit dependency between earlier and later substeps. In that sense, it extends the logic of CoT without yet moving to the kind of branching exploration that appears in later search-based methods.

A related development was the use of reasoning traces not only as prompts

but also as material for improving the model itself. STaR showed that a model can generate its own rationales, retain the ones that lead to correct answers and use them to further train its reasoning behavior [19]. This is an important shift because it suggests that progress in reasoning does not come only from better prompting at inference time but also from better ways of collecting and reusing intermediate traces. At the same time, later work raised doubts about how far these gains should be taken as evidence of genuine deliberative reasoning. Some studies suggest that CoT is especially effective for mathematics and symbolic reasoning but much less consistently helpful outside those settings [35]. Others show that models often fail to reliably correct their won reasoning without external feedback and that resulting explanations are not always faithful accounts of how a decision was actually made [36, 20]. Together these limitations help explain why later research increasingly turned toward methods that treat reasoning not simply as single generated chain but as a process that may require evaluation and search over alternatives.

4 Evaluating Deliberative Reasoning Claims

The inference time methods discussed above make modern language models look increasingly deliberative. Rather than relying on a single left-to-right chain, recent systems can sample alternatives, decompose problems into substeps, evaluate partial solutions, backtrack from failed paths and in some cases combine reasoning with explicit planning or action. Yet these same developments also make the central interpretive question harder. If performance improves through search, sampling, rationale reuse, external evaluation or additional test-time computation, then it becomes less obvious what portion of that improvement should be attributed to deliberative reasoning by the model itself. This raises a more demanding question for the literature: what kind of evidence is actually sufficient to justify a claim of "deliberative reasoning" rather than simply more effective inference-time problem solving [24, 37, 38]? A model may do better on hard benchmarks, produce longer intermediate outputs or seem to compare alternatives before answering and yet none of this by itself tells us very much about the underlying process that generated the result [39, 40, 41]. If the aim is to evaluate claims about deliberations rather than just record performance gains, then the evidential bar has to be set higher.

Part of what makes these claims feel more plausible today is that empirical picture has genuinely changed. Early work on CoT prompting showed that sufficiently large language models could imporve sharply on arithmetic, common sense and symbolic reasoning tasks when prompted to generate intermediate steps [15]. Subsequent methods extended this basic idea in several directions. STaR used self generated rationales to bootstrap better reasoning behavior [19]. ReAct combined reasoning with action [31]. Tree of Thoughts introduced explicit search over candidate reasoning paths instead of relying on a single linear chain [22]. More recent approaches move beyond prompt design alone. DeepSeek-R1 present reinforcement learning as a way to induce

long form reasoning behavior directly, while s1 argues that careful control of test-time compute can recover meaningful scaling effects even from relatively small amount of curated data [42, 43]. Meta CoT makes a related but stronger point arguing that standard CoT often captures only a cleaned up final trace rather than the fuller process of exploration, checking and revision that difficult reasoning may involve [44]. Taken together, these developments make it understandable why stronger claims about deliberation now seem more credible than they did in the earlier prompting era.

Even so, behavioral improvement does not settle the issue. Better performance with intermediate steps is not the same as evidence that those steps reflect a genuine internal deliberative process. One reason for caution comes from the CoT literature itself. Wang et al. show that even "invalid" reasoning demonstrations can preserve much of the benefit of valid ones, provided they stay relevant to the task and maintain a sensible ordering [39]. Related work on reasoning-step length reaches a similar conclusion. Longer traces can improve performance even when they introduce little genuinely new information and even incorrect rationales may still help if they preserve enough inferential depth [40]. More broadly, prompt based decomposition methods such as Least-to-Most, self-consistency decoding and self-verification suggest that performance can improve when computation is restructured or sample differently at inference time, without that resolving whether the resulting trace is a faithful account of the model's reasoning process [18, 34, 45]. Findings like these make it hard to move cleanly from "the model produced a chain" to "the model deliberated in the way that chain describes."

This is exactly why faithfulness becomes such a central issue. When a model produces a polished multi-step explanation, does that explanation actually reflect the process that caused that answer? Turpin et al. give strong reasons for skepticism, showing that language models can produce explanations that rationalize answers shaped by hidden prompt biases while never mentioning those biases at all [37]. Lanham et al. arrive at a similar concern from a different direction. By truncating, paraphrasing, perturbing or replacing stated reasoning, they find that faithfulness varies widely across tasks and that larger, more capable models are often less faithful on many of the tasks they examine. Figure 3 illustrates the main intervention types used in that analysis, inlcuding early answering, adding mistakes, paraphrasing and filler-token replacement [38]. That result matters because it cuts against a tempting assumption in the literature. Better models are not necessarily more transparent ones. In response to this problem, Faithful CoT proposes routing reasoning through a symbolic intermediate representation and deterministic solver, tightening the connection between intermediate steps and final answers [46]. Seen in this light, work on faithful reasoning is not just a refinement of CoT. It is also an implicit admission that ordinary free-form reasoning traces are not enough on their own, to support a strong claim about deliberation.

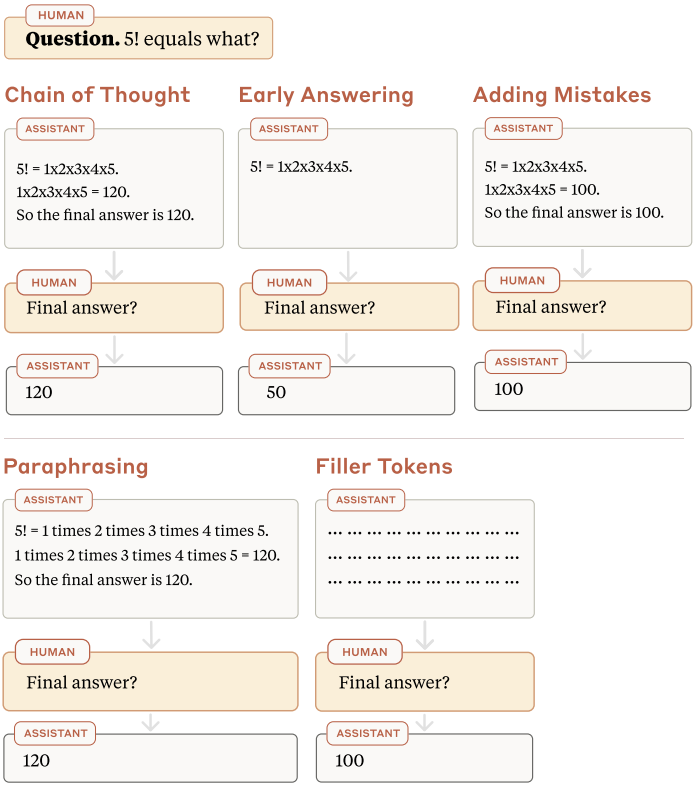

Figure 3: Faithfulness interventions used to test whether a model's final answer depends on its state CoT. Adapted form [38].

A seperate concern is robustness. Even setting faithfulness aside, a genuinely deliberative process should not fail whenever the surface form of a problem changes in a way that is logically irrelevant. Yet several papers suggest that current systems remain brittle in exactly this sense. Chen et al. show that reordering logically equivalent premises can substantially reduce performance even though the logical content of the problem is unchanged [47]. Shi et al. show that irrelevant context can significantly distract models on arithmetic reasoning tasks [48]. Jian et al. argue that apparent reasoning success may still depend heavily on superficial token-level regularities rather than robust abstraction over logical structure [41]. Recent work on stable reasoning adds another layer to this critique by arguing that metrics such as Pass@k can overstate reasoning ability since they measure potential under repeated sampling more than consistency on an ordinary attempt [49]. On this view, a system that succeeds mainly under favorable phrasing, repeated retries or benchmark-specific regularities has not yet earned a strong claim to deliberative reasoning.

The evaluation problem makes the picture even more complicated. If public

reasoning benchmarks are increasingly entangled with the training data ecosystem, then strong performance on static evaluations becomes harder to interpret as evidence of deeper reasoning ability [50]. This has motivated a move toward dynamic evaluation where new or transformed problems are generated in order to reduce contamination and make it more difficult to explain benchmark success through memorization or overlap [50, 51]. That does not completely resolve the issue. As recent surveys note, dynamic evaluation still lacks fully settled standards of its own [50]. Still, it reinforces an important point that can easily be overlooked: benchmark gain are not neutral evidence. Their meaning depends in part on the quality, freshness and construction of the benchmark itself.

At this stage, it would be easy to wing too far in the opposite direction and conclude that claims of deliberative reasoning are mostly illusory. That would be too strong as well. Mechanistic and representational studies do provide at least some evidence that nontrivial internal structure emerges in reasoning settings. Cabannes et al. for instance, identify specialized attention mechanisms associated with iterative reasoning in controlled transformer tasks which suggests that CoT behavior can sometimes correspond to real internal computational organization rather than simple verbosity [52]. Yang et al. likewise find evidence that models can recover intermediate bridge entities in latent multi-hp reasoning, although the results are uneven and much stronger for the first hop than for full multi-hop traversal [53]. These studies do not show that current language models possess a clean and general form of deliberative reasoning. But they do matter because they suggest that the right conclusion is not that all reasoning claims are empty or merely stylistic. The evidence is better understood as partial, uneven and highly dependent on what exactly is being measured.

For that reason, the most defensible position is a demanding but balanced one. Improved performance, longer CoT and RL-trained reasoning traces should be treated as "suggestive evidence", not decisive proof of deliberative reasoning [43, 42, 44]. Stronger claims require convergence across several dimensions at once: behavioral improvement, faithfulness of intermediate steps, robustness under logically irrelevant perturbations and evaluation protocols that are resistant to contamination and metric inflation [37, 38, 41, 49, 50]. From this perspective, recent reasoning-oriented models are geuniely important because they show that post-training methods and inference-time compute can unlock capabilites that earlier prompting methods only exposed imperfectly [42, 43]. At the same time, the current literature still falls short of justifying a straightforward equation between benchmark gains and transparent, stable, deliberative inference. The stronger claim may yet turn out to be correct but it has to earned under a stricter evidential standard than performance gains alone can satisfy.

5 Limitations of Current Architectures

Current LLMs often imitate causal discourse convincingly yet remain structurally misaligned with formal causal inference, especially interventions and

counterfactuals. Empirical causal benchmarks that explicitly test beyond associational reasoning find (i) systematic performance degradation from association to intervention to counterfactual queries [54] and (ii) brittle out-of-distribution (OOD) generalization under variable renaming or paraphrasing [55].

Scaling and inference-time scaffolds (prompting, retrieval, search) can raise benchmark scores but do not, by themselves, establish that the architecture implements causal operators. Mitigations under active study such as causal pretraining, interventional/counterfactual fine-tuning, hybrid neuro-symbolic pipelines, and counterfactual retrieval, show promise but introduce new compute/data/tooling trade-offs and remain incomplete as general-purpose solutions [56] [57] [58].

A core architectural mismatch is that autoregressive "causal masking" enforces temporal precedence in token prediction, not causal semantics in Pearl's sense. CausalProbe-2024 explicitly argues that the autoregression mechanism of transformer LLMs is "not inherently causal" for causal inference, and demonstrates large drops on freshness-controlled causal Q&A constructed after purported training cutoffs, consistent with shallow associative behavior rather than intervention-level reasoning [59].

LLM internal representations are distributed and often entangled with lexical form. This can inhibit variable-based abstraction needed for do-calculus and counterfactual reasoning. Evidence from Corr2Cause's OOD perturbations is consistent with surface-level anchoring rather than symbolic-variable manipulation [55] .

Reasoning traces add a distinct limitation: chain-of-thought (CoT) may be brittle and unfaithful. Turpin et al. show CoT explanations can systematically misrepresent what drives predictions: adding biasing prompt features (e.g., option reordering) can drop accuracy by up to 36% while explanations omit the true driver [37]. Lanham et al. find models vary in how much they condition on CoT; as models scale, faithfulness can decrease on many tasks, and interventions on the CoT reveal instability [38]. These behaviors undermine using CoT as evidence of internal causal reasoning rather than post-hoc rationalization.

Future progress toward reliable, causal artificial intelligence will necessitate a paradigm shift away from purely generative token prediction toward explicit neuro-symbolic integration, representation-based world models, and architectures capable of transparent, verifiable, and goal-oriented planning. Until such fundamental architectural transformations are achieved, LLMs must be treated as powerful semantic interfaces rather than autonomous cognitive reasoning engines.

6 Trust and Deployment Risks

Currently, LLMs are seen as a potential tool to revolutionize many industries like finance, healthcare, and defense. As such, many have looked into how to apply them.

6.1 Finance

In finance, much interest has been focused on sentiment analysis, information extraction, question answering, and stock movement prediction [60]. LLMs appear to be particularly effective for certain tasks like sentiment analysis and information extraction, with fine-tuned models achieving F1 scores above 0.8, and even base frontier models performing competitively in zero-shot and few-shot settings [61].

For question answering, frontier models are approaching human-level performance due to advances in pretraining techniques and prompting strategies such as Chain-of-Thought reasoning [62]. However, performance remains uneven, especially in tasks requiring quantitative reasoning over financial data. For example, models fine-tuned for finance still underperform general-purpose LLMs like GPT-4 on numerical reasoning benchmarks, highlighting a key limitation in deployment for high-stakes financial decision-making [61].

Despite these promising capabilities, several deployment risks remain. First, LLMs demonstrate limited reliability in financial prediction tasks such as stock movement forecasting. Even domain-specific models struggle to outperform simple baselines, suggesting that real-world predictive deployment remains highly uncertain [61]. Second, hallucination and overconfidence pose serious trust issues, particularly in domains like finance where incorrect outputs may lead to significant economic consequences. Third, domain-specific nuances and rapidly changing market conditions can degrade model performance over time, raising concerns about robustness and generalization.

Finally, there are broader systemic risks. The use of LLMs in finance may amplify misinformation and enable manipulation of markets through automated content generation. Since there is currently no reliable way to inspect an LLM's thinking process and potentially steer it, these concerns are the main obstacles against deployment.

6.2 Healthcare

In healthcare, LLMs are being used for sentence classification, clinical information extraction, and QA very similarly to finance. For sentence classification, recent systematic reviews show that LLMs are increasingly applied to categorize clinical notes, patient communications, and research abstracts for tasks such as diagnosis coding, triage, and surveillance, often outperforming traditional machine learning models on benchmark datasets[63]. However, these reviews also emphasize that most studies are conducted in curated research corpora, rely on precision metrics rather than patient safety outcomes, and rarely evaluate performance under realistic distribution changes, leaving open questions about how robust such classifiers are when deployed in live clinical workflows [63]. In addition, the performance of an LLM on these benchmarks significantly degrade when you shuffle the answer choices, remove the image, or include a distractor which highlights the lack of robustness of these benchmarks as well as of these models [64]

Furthermore, a systematic review found only 4 empirical deployment studies between 2024 and 2025, all based on GPT-family models integrated into outpatient communication, mental health support, inbox message drafting, and clinical data extraction; while these deployments reported gains in operational efficiency, user satisfaction, and reduced workload, they also highlighted performance variability across data types, limited generalizability, regulatory delays, and the absence of robust post-deployment monitoring and standardized outcome metrics, underscoring the need for multi-site validation, human oversight, and implementation frameworks tailored to clinical settings [65].

6.3 Defense

Large language models are beginning to be explored for intelligence analysis, operational planning, and decision support in defense settings, but emerging evidence from wargaming and early field studies highlights serious concerns about escalation risk, reliability, data security, and trustworthiness that must be addressed before deployment in high-stakes military workflows. For example, one study testing frontier models on simulated war games found that they chose nuclear signaling in 95% of simulated crises, a clear action that goes against human values [66]. This shows that models are not well aligned with human values in this space. In addition, studies have found that LLMs suffer from issues with data leakage, effective finetuning, and poor reproducibility and assessment [67]. In addition, they still have biases and hallucinations at runtime and are prone to jailbreaks making them too unreliable [68]. Much of this unreliability comes from the fact that generative models like LLMs are non-deterministic which fundamentally differs with what is needed in weapons control [69].

7 Conclusion

In summary, this review examined whether the impressive performance of modern large language models reflects genuine deliberative reasoning or sophisticated statistical pattern matching. While techniques such as chain-of-thought prompting, structured search, and inference-time computation improve performance on complex tasks, the evidence suggests that these gains often arise from sampling, scaffolding, and external evaluation rather than internally faithful reasoning. Current transformer architectures remain limited in causal understanding, counterfactual reasoning, and robust out-of-distribution generalization, and reasoning traces can be brittle or unfaithful to the model's actual decision process. These limitations raise concerns about interpretability and reliability, particularly in high-stakes domains such as finance, healthcare, and defense. Moving forward, advancing toward trustworthy reasoning systems will likely require architectural innovations, causal learning frameworks, and stronger evaluation methodologies that distinguish true deliberation from improved inference strategies. Until then, LLMs should be viewed as powerful but fundamentally constrained tools whose reasoning-like behavior does not yet

constitute genuine understanding.

References

[1] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950. Accessed at https://courses.cs.umbc.edu/471/papers/turing.pdf.

[2] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. The dartmouth summer research project on artificial intelligence. In *Proceedings of the Dartmouth Summer Research Project on Artificial Intelligence*. Dartmouth College, 1956. Accessed at http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

[3] Rudolf Carnap. *Introduction to symbolic logic and its applications*. Courier Corporation, 2012.

[4] Clarence Irving Lewis, Cooper Harold Langford, and P Lamprecht. *Symbolic logic*, volume 170. Dover publications New York, 1959.

[5] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[6] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[7] Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, NIPS'89, page 396–404, Cambridge, MA, USA, 1989. MIT Press.

[8] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[9] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger,

Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[11] OpenAI. Chatgpt: Optimizing language models for dialogue, November 2022. Accessed at https://openai.com/blog/chatgpt.

[12] Leonardo De Angelis, Francesco Baglivo, Guglielmo Nardi, Paolo Torino, Paolo Ferragina, and Klaus Panetta. Chatgpt and the rise of large language models: The new ai-driven infodemic threat in public health. *Frontiers in Public Health*, 11:1166120, 2023.

[13] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

[14] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, David Grangier, Eric Horvitz, Ece Kaur, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4, 2023. Microsoft Research preprint.

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichien, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2023.

[16] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[17] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. *ArXiv preprint*, abs/2210.00720, 2022.

[18] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[19] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[20] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in

chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[21] Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. Chain-of-thought is not explainability, July 2025. Preprint.

[22] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[23] Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. A tree-of-thoughts to broaden multi-step reasoning across languages. *arXiv preprint arXiv:2311.08097*, 2023.

[24] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhijiang Guo, Le Song, and Cheng-Lin Liu. From system 1 to system 2: A survey of reasoning large language models, 2025.

[25] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

[26] Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. Forest-of-thought: Scaling test-time compute for enhancing llm reasoning, 2025.

[27] Jiahao Qiu, Yifu Lu, Yifan Zeng, Jiacheng Guo, Jiayi Geng, Chenhao Zhu, Xinzhe Juan, Ling Yang, Huazheng Wang, Kaixuan Huang, Yue Wu, and Mengdi Wang. Treebon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling, 2025.

[28] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making large language models better reasoners with step-aware verifier, 2023.

[29] Zhuohao Yu, Weizheng Gu, Yidong Wang, Xingru Jiang, Zhengran Zeng, Jindong Wang, Wei Ye, and Shikun Zhang. Reasoning through execution: Unifying process and outcome rewards for code generation, 2025.

[30] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms, 2024.

[31] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

[32] Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

[33] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*, 2024.

[34] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023.

[35] Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning, 2025.

[36] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet, 2024.

[37] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

[38] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023.

[39] Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters, 2023.

[40] Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. The impact of reasoning step length on large language models, 2024.

[41] Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J. Su, Camillo J. Taylor, and Dan Roth. A peek into token bias: Large language models are not yet genuine reasoners, 2024.

[42] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, September 2025.

[43] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès,

and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025.

[44] Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, Louis Castricato, Jan-Philipp Franken, Nick Haber, and Chelsea Finn. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought, 2025.

[45] Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification, 2023.

[46] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning, 2023.

[47] Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models, 2024.

[48] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context, 2023.

[49] Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei Zhang, Songyang Zhang, and Kai Chen. Are your llms capable of stable reasoning?, 2025.

[50] Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, and Baishakhi Ray. Recent advances in large langauge model benchmarks against data contamination: From static to dynamic evaluation, 2025.

[51] Simin Chen, Pranav Pusarla, and Baishakhi Ray. Dynamic benchmarking of reasoning capabilities in code large language models under data contamination, 2025.

[52] Vivien Cabannes, Charles Arnal, Wassim Bouaziz, Alice Yang, Francois Charton, and Julia Kempe. Iteration head: A mechanistic study of chain-of-thought, 2024.

[53] Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning?, 2025.

[54] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models, 2024.

[55] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation?, 2024.

[56] Aniket Vashishtha, Qirun Dai, Hongyuan Mei, Amit Sharma, Chenhao Tan, and Hao Peng. Executable counterfactuals: Improving llms' causal reasoning through code, 2025.

[57] Huaiyu Qin, Chunyu Wei, Yueguo Chen, and Yunhai Wang. Counterfactual reasoning for retrieval-augmented generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[58] Sen Yang, Xin Li, Leyang Cui, Lidong Bing, and Wai Lam. Neuro-symbolic integration brings causal and reliable reasoning proofs. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5747–5759, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[59] Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. Unveiling causal reasoning in large language models: Reality or mirage?, 2025.

[60] Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. A survey on large language models for critical societal domains: Finance, healthcare, and law, 2024.

[61] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance, 2023.

[62] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[63] Hajar Sakai and Sarah S Lam. Large language models for health care text classification: Systematic review. *JMIR AI*, 5:e79202, February 2026.

[64] Yu Gu, Jingjing Fu, Xiaodong Liu, Jeya Maria Jose Valanarasu, Noel CF Codella, Reuben Tan, Qianchu Liu, Ying Jin, Sheng Zhang, Jinyu Wang, Rui Wang, Lei Song, Guanghui Qin, Naoto Usuyama, Cliff Wong, Hao Cheng, HoHin Lee, Praneeth Sanapathi, Sarah Hilado, Tristan Naumann, Javier Alvarez-Valle, Jiang Bian, Mu Wei, Khalil Malik, Lidong Zhou, Jianfeng Gao, Eric Horvitz, Matthew P. Lungren, Doug Burger, Eric Topol, Hoifung Poon, and Paul Vozila. The illusion of readiness in health ai, 2025.

[65] Yaara Artsi, Vera Sorin, Benjamin S. Glicksberg, Panagiotis Korfiatis, Girish N. Nadkarni, and Eyal Klang. Large language models in real-world clinical workflows: a systematic review of applications and implementation. *Frontiers in Digital Health*, 7:1659134, 2025.

[66] Kenneth Payne. Ai arms and influence: Frontier models exhibit sophisticated reasoning in simulated nuclear crises, 2026.

[67] Shannon K. Gallagher, Jasmine Ratchford, Tyler Brooks, Bryan P. Brown, Eric Heim, William R. Nichols, Scott Mcmillan, Swati Rallapalli, Carol J. Smith, Nathan Vanhoudnos, and et al. Assessing llms for high stakes applications. *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*, page 103–105, Apr 2024.

[68] William N. Caballero and Phillip R. Jenkins. On large language models in national security applications, 2024.

[69] Mary Cummings. Prohibiting generative AI in any form of weapon control. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025.