Opening the Black Box: A Survey on the Mechanisms of Multi-Step Reasoning in Large Language Models

Liangming Pan[1†], **Jason Liang**[2], **Jiaran Ye**[3],
Minglai Yang[3,4], **Xinyuan Lu**[5], **Fengbin Zhu**[5†]

[1]Peking University [2]Stanford University [3]Tsinghua University
[4]University of Arizona [5]National University of Singapore

liangmingpan@pku.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated remarkable abilities to solve problems requiring multiple reasoning steps, yet the internal mechanisms enabling such capabilities remain elusive. Unlike existing surveys that primarily focus on engineering methods to enhance performance, this survey provides a comprehensive overview of the *mechanisms* underlying LLM multi-step reasoning. We organize the survey around a conceptual framework comprising seven interconnected research questions from how LLMs execute *implicit multi-hop reasoning* within hidden activations to how *verbalized explicit reasoning* remodels the internal computation. Finally, we highlight five research directions for future mechanistic studies.

1 Introduction

Large Language Models (LLMs) have demonstrated an impressive ability to carry out ***multi-step reasoning***, which involves the process of drawing conclusions through a sequence of intermediate steps, where each step builds on the previous one. Multi-step reasoning has been widely regarded as one of the most fundamental forms of reasoning (Hou et al., 2023a; Guo et al., 2025). It serves as the backbone of advanced tasks such as deep question answering, mathematical problem solving, logical deduction, code generation, and planning (Chen et al., 2021; Wei et al., 2022b; OpenAI, 2023; Yang et al., 2024a; Dubey et al., 2024; Guo et al., 2024; DeepSeek-AI et al., 2025).

Multi-step reasoning in LLMs generally takes on two distinct forms. ***Implicit reasoning*** involves performing multi-hop inference entirely within the model's hidden activations, delivering a correct final answer without verbalizing intermediate steps. In contrast, ***explicit reasoning***, exemplified by *Chain-of-Thought* (CoT) (Wei et al., 2022b), instructs the model to externalize the reasoning process into a sequence of natural language tokens. Remarkably, modern LLMs have exhibited strong performance in both paradigms (Chu et al., 2024a; Chen et al., 2025a; Li et al., 2025b). Building on this empirical success, the ***internal mechanisms*** that enable such capabilities become scientifically intriguing. For implicit reasoning, a key puzzle is *how* multi-step reasoning capabilities emerge from simple next-token prediction training, and *how* LLMs internally carry out multi-step computations. For explicit CoT reasoning, critical questions persist about *why* CoT can elicit superior reasoning capabilities and *whether* the generated rationale faithfully reflects the model's actual decision-making process. Understanding these mechanisms is not only a matter of scientific curiosity but also a prerequisite for building more reliable, controllable, and human-aligned reasoning systems.

Although we still lack a unified mechanistic theory, a growing body of literature seeks to *open the black box of LLM multi-step reasoning* and has made significant progress. In this paper, we aim to provide a comprehensive overview of these works. Unlike existing surveys (Huang and Chang, 2023; Chu et al., 2024b; Chen et al., 2025a) that primarily focus on *enhancing* reasoning (*e.g.*, through tool use, retrieval augmentation, or self-correction), our survey explicitly focuses on ***understanding mechanisms***, a perspective that has been largely overlooked in previous reviews. As illustrated in Figure 1, we identify seven pivotal, interconnected, and progressive *research questions (RQs)* to form the cognitive framework of our survey. These questions form a cohesive narrative, covering analytical methods and key findings from the hidden internal dynamics of latent reasoning to the visible mechanisms of explicit CoT reasoning. We end by pointing out five open research questions that remain under-explored but are essential for the future roadmap of mechanistic understanding.

†Corresponding Authors

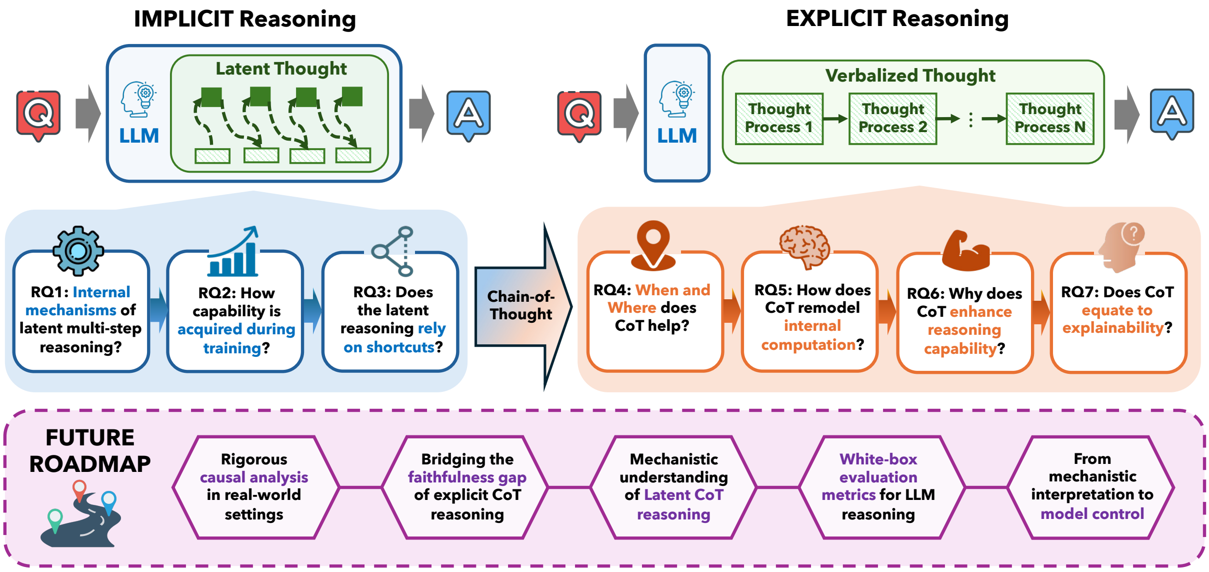

Figure 1: The cognitive framework and organizational structure of this survey. We explore the mechanisms of multi-step reasoning through two distinct paradigms: *Implicit Reasoning* and *Explicit Reasoning*, through seven interconnected *Research Questions*. The bottom panel highlights five strategic directions for future research.

2 Implicit Multi-Step Reasoning

Multi-hop implicit reasoning is the process of answering a question by combining multiple pieces of information across several steps. Unlike explicit reasoning, the intermediate links are not directly stated and must be inferred using background knowledge or context. Mechanistic study of multi-hop implicit reasoning is important because it reveals whether models truly perform step-by-step reasoning or rely on shallow shortcuts. Such understanding improves interpretability and trust in LLMs, and it guides the development of models that generalize more reliably.

2.1 What are the internal mechanisms of latent multi-step reasoning?

Recent mechanistic studies have begun to unveil how LLMs carry out latent multi-hop computation entirely in their hidden states (Yang et al., 2024b; Biran et al., 2024; Brinkmann et al., 2024a). These studies employing causal probing, mechanistic tracing, and representational analysis have collectively revealed a *staged* internal process in which intermediate results are computed and transformed layer by layer, ultimately contributing to the final output. In essence, transformers appear to implement an internal chain-of-thought spread across their depth.

Functional specialization of layers. A major body of work explores *layer specialization*, aiming to identify the distinct computational roles each layer plays during multi-hop inference. Using *Patchscopes* (Ghandeharioun et al., 2024) together with a novel intervention technique termed *back-patching*, Biran et al. (2024) uncovered a sequential computation pathway in which early layers identify the bridge entity, which is then propagated forward and exploited by later layers to complete the inference. Complementarily, Li et al. (2024a) applied logit lens analysis (nostalgebraist, 2020) and found that implicit reasoning representations emerge in intermediate layers and have a causal influence on generating the final answer. Extending this perspective, Yu et al. (2025) traced logits through the network via a neuron-level logit flow method and observed that even a single-hop query is solved in multiple distinct stages—entity subject enrichment, entity attribute extraction, relation subject enrichment, and relation attribute extraction—each of which is localized to different layers. More recently, Yang et al. (2025d) showed that this layer-wise reasoning also applies at the task level: for composite instructions, models execute different subtasks at different depths, forming a staged computation across layers. All the above studies provided evidence of functional specialization of transformer layers in multi-hop reasoning.

Uncovering fine-grained reasoning structures. Beyond layer specification, another line of work aims to recover more fine-grained implicit reasoning structures from model internals. *MechanisticProbe* (Hou et al., 2023b) introduced an attention-probing technique to extract latent reasoning trees from transformer activations. They showed that on synthetic and natural tasks with GPT-2 and LLaMA, models often perform procedural reasoning layer by layer, with lower layers selecting statements and higher layers executing reasoning steps. Complementing these findings, Brinkmann et al. (2024b) analyzed a small transformer trained on a symbolic tree path-finding task, finding that it implements a backward chaining algorithm: deduction heads climb trees one level per layer, register tokens act as working memory for parallel subpaths, and a one-step lookahead heuristic compensates when chaining is insufficient. Together, these studies demonstrate that transformers can adopt structured, algorithm-like reasoning strategies beyond memorization, albeit within the limits of the model's depth (to be discussed below).

Layer depth as the primary bottleneck for implicit reasoning. Theoretical and empirical studies indicate that the number of reasoning steps a model can perform implicitly is strictly limited by its depth. Merrill and Sabharwal (2024) theoretically demonstrated that a standard Transformer with constant depth cannot solve inherently serial problems that require computation scaling with input size, *e.g.*, parity or graph connectivity. In practice, Yu (2025) and Guo et al. (2025) found that specific multi-hop reasoning tasks require a minimum threshold of layers to resolve; if a model is too shallow, the "latent chain" is cut short, and the reasoning fails. Saunshi et al. (2025) formally established that an L-layer Transformer can simulate an m-step explicit reasoning process, provided L is sufficiently large to accommodate the iterative forward passes required. All these works revealed a close correlation between layer depth and the implicit reasoning capabilities of the model.

Why implicit reasoning sometimes fails. Identifying how and why implicit reasoning sometimes fails has also been illuminating. Biran et al. (2024) discovered that many failures stem from delayed resolution of the first hop, and showed that rerunning computations via back-patching can correct these errors. Li et al. (2024a) found that failures frequently arise from the improper generation or utilization of implicit reasoning results. To address this, they proposed CREME, a lightweight model-editing technique that patches specific multi-head self-attention modules, leading to improved compositional reasoning generalization with minimal disruption to unrelated predictions. In the context of two-hop queries ("e_1's r_1's r_2 is e_3"), Yu et al. (2025) showed that errors often occur when high-layer features at the r_1 position overemphasize the intermediate entity e_2, outweighing the logits for the correct final entity e_3. This finding revealed that LLMs internally build and combine entity–relation representations in a staged manner, but positional interference can derail multi-hop reasoning. To fix this, they introduced a back-attention mechanism allowing lower layers to reuse higher-layer information from other positions, which substantially improved multi-hop accuracy. However, even with such interventions, certain transformers still struggle to reliably chain more than one reasoning step. For example, Yang et al. (2024b) found that *LLaMA-2* models, while reliably recalling a needed bridge entity, often fail to apply it to the second hop, highlighting limits in architecture that impede consistent multi-step chaining.

Takeaway: Implicit multi-hop reasoning in LLMs is not a monolithic capability but rather an orchestrated, layered process, with distinct modules and pathways specializing in different phases of the reasoning chain. For example, probing and intervention studies showed that intermediate results, *e.g.*, bridge entities, are computed and passed along inside the network. Nevertheless, such implicit reasoning is constrained by the inherent architecture of transformers, for example, their fixed depth.

2.2 How latent multi-step reasoning capability is acquired during training?

Models do not possess latent reasoning capabilities at initialization. If multi-hop reasoning is implemented via specialized internal circuits discussed in Section 2.1, a critical question arises: *how do these circuits emerge in the first place?* Research into training dynamics reveals that implicit reasoning is an *acquired* behavior that emerges during the training process through distinct phase transitions.

Grokking marks the shift from memorization to reasoning. Recent studies (Wang et al., 2024; Ye et al., 2025; Zhang et al., 2025c; Abramov et al., 2025) suggested that LLMs do not learn multi-step reasoning gradually; instead, they often undergo *phase transitions* during training where reasoning capabilities appear *suddenly* rather than continuously. In other words, a model might spend many updates seemingly memorizing or floundering, then "grok" the underlying reasoning algorithm after a certain point. This phenomenon, known as *"grokking"*, was initially observed in deep networks trained on other tasks such as modular arithmetic, where generalization performance spikes long after training accuracy has saturated (Power et al., 2022; Olsson et al., 2022; Wei et al., 2022a).

In the context of multi-hop implicit reasoning, this phenomenon of transformers transitioning from early-stage *memorization* to later-stage *generalization* was first observed by Wang et al. (2024) through training transformers from scratch on symbolic reasoning tasks. They found that the multi-hop reasoning capability emerges only through *grokking*, where an early *memorizing circuit* is gradually replaced by a more efficient *generalizing circuit* due to optimization bias and weight decay. Ye et al. (2025) corroborated this phase transition, proposing a *three-stage trajectory*: (i) rapid memorization, (ii) delayed in-distribution generalization, and (iii) slower cross-distribution generalization, with persistent OOD bottlenecks at the second hop. Mechanistically, they employed *cross-query semantic patching* to localize the "bridge" entity and a *cosine-based representational lens* to reveal that generalization coincides with *mid-layer clustering* of intermediate entity representations.

Factors influencing the emergence of reasoning. The transition from memorization to generalization is not random; studies revealed that it is governed by specific properties. One of the primary determinants is the *training data distribution*. Wang et al. (2024) demonstrated that the speed of grokking correlates strongly with the *ratio of inferred to atomic facts* ϕ in training. A higher ratio of compositional examples forces the model to abandon inefficient memorization in favor of the generalizing circuit. Expanding this to real-world scenarios, Abramov et al. (2025) found that natural corpora often lack sufficient connectivity (low ϕ) to trigger this transition, but data augmentation with synthetic inferred facts can artificially raise ϕ above the critical threshold required for circuit formation. Beyond data distribution, the *scale of the training data* also matters. Yao et al. (2025b) revealed a scaling law: the data budget required to learn implicit k-hop reasoning grows exponentially with k, though curriculum learning can significantly mitigate this cost. From an optimization perspective, Zhang et al. (2025c) identified *complexity control* parameters as crucial factors. They found that smaller initialization scales and stronger weight decay bias the optimization process toward low-complexity, rule-like solutions rather than high-complexity, memory-based mappings, thereby accelerating the emergence of reasoning capabilities. Finally, Li et al. (2025c) observed that in large-scale pretraining, grokking is *asynchronous and local*; different domains and data groups undergo this memorization-to-generalization transition at different times depending on their inherent difficulty and distribution heterogeneity.

Takeaway: Implicit multi-hop reasoning capability is an *acquired* capability that emerges via *grokking*—a phase transition from surface-level memorization to structured reasoning. This transition is not automatic; it is governed by critical factors, including the training data distribution, the data scale, and complexity control via optimization biases.

2.3 To what extent does multi-step reasoning rely on shortcuts?

While the training dynamics discussed in § 2.1 suggest that structured reasoning circuits can emerge, growing mechanistic evidence has also uncovered a more complex and often discouraging reality regarding model internals. Models frequently bypass genuine multi-step reasoning, relying instead on *"shortcuts"*—statistical correlations or surface-level heuristics that mimic reasoning without performing the underlying computation (Kang and Choi, 2023; Elazar et al., 2024; Yang et al., 2025b).

Factual shortcuts bypass intermediate reasoning. A primary form of shortcutting involves exploiting direct associations between the subject and the final answer, effectively skipping the intermediate steps. Ju et al. (2024) investigated this in the context of knowledge editing, finding that failures often stem from "shortcut neurons" that encode a direct link between the first and last entities, ignoring the multi-hop structure. Mechanistically, Yang et al. (2025c) used *Patchscopes* (Ghandeharioun et al., 2024) to distinguish valid reasoning from shortcuts. They observed that genuine implicit reasoning coincides with the model constructing a hidden representation of the intermediate bridge entity. In contrast, shortcut-prone queries bypass this internal construction entirely. When these direct shortcuts are removed, model performance drops by nearly a factor of three, revealing that much of the perceived reasoning capability is illusory.

Shortcuts based on surface-level pattern matching. Beyond factual associations, models also latch onto structural regularities in the training data. Lin et al. (2025) analyzed implicit arithmetic reasoning and found that models often adopt a "bag-of-words" heuristic, treating operations as commutative even when they are not. While this shortcut works for fixed-template examples, performance collapses when premise order is randomized, proving the model had not learned the robust sequential logic. Similarly, Guo et al. (2025) found that in the presence of context distractors, pretrained models default to a heuristic of guessing based on surface plausibility. However, they also noted a positive trajectory: fine-tuning can force a phase transition where the model shifts from this shallow guessing behavior to a sequential query mechanism that explicitly retrieves intermediate entities.

Takeaway: LLMs frequently bypass the "latent reasoning chain" via *factual shortcuts* (direct input-output associations) or *structural heuristics* (exploiting surface patterns like commutativity). This underscores the need for shortcut-free evaluation protocols and training setups that force models to construct and reuse intermediate representations.

3 Explicit Multi-Step Reasoning

Implicit reasoning operates entirely within the fixed computational budget of the model's hidden states; therefore, it is bounded by the depth bottleneck and frequently falls prey to shortcuts. *Explicit multi-step reasoning* fundamentally alters this paradigm. By prompting an LLM to produce a step-by-step *Chain-of-Thought* (CoT), the reasoning process is externalized into a sequence of natural language tokens, effectively extending the computational capacity beyond the model's layers. CoT has been shown to unlock significantly better performance on tasks that require reasoning. In this section, we dissect the mechanisms of this paradigm through four progressive research questions (§ 3.1-§ 3.4).

3.1 Where and When Does CoT Help?

On which tasks does CoT help? To uncover this, Sprague et al. (2025) conducted a large-scale meta-analysis across 20 benchmarks and found that prompting with CoT yields large gains primarily on *math and symbolic logic tasks*, with far smaller or even negative gains on other domains. Suzgun et al. (2023) similarly showed that many *BIG-Bench Hard* tasks (Srivastava et al., 2023), which had stumped standard few-shot prompts, become solvable with CoT. These were precisely tasks requiring multi-step reasoning, *e.g.*, symbolic manipulation, compositional logic. However, for knowledge-heavy tasks like MMLU (Hendrycks et al., 2021) or commonsense reasoning, CoT often provides negligible improvement (Sprague et al., 2025). In certain cases, CoT can even degrade accuracy. For example, Liu et al. (2024) examined cognitive-psychology tasks where additional deliberation harms human performance, *e.g.*, certain trick riddles or intuitive judgment problems. They found that CoT substantially degraded accuracy on such tasks, and it tends to distract the model into over-complicating a problem that might have been solved via intuition. A complementary study on Blocksworld planning (Stechly et al., 2024) found that CoT helps only when the prompt examples closely match the test distribution, and the gains quickly deteriorate if the test problem's complexity exceeds that seen in the exemplars.

What factors influence the efficacy of CoT? Beyond task-level evaluations, empirical studies have shown that CoT performance can be dramatically influenced by many features of the CoT prompt. First, studies (Ye and Durrett, 2022; Madaan et al., 2023; Wang et al., 2023) reveal that the *relevance and ordering* of exemplars matter more than their semantics; models can still derive correct answers from invalid rationales if the prompt maintains a coherent structure. Second, the length of reasoning is another critical factor, with Jin et al. (2024) identifying that the number of reasoning steps significantly modulates model performance. Finally, CoT is surprisingly sensitive to phrasing; minor input perturbations can substantially bias models' answers (Turpin et al., 2023; Sadr et al., 2025).

Why do these factors influence CoT efficacy? To explain the mechanisms underlying these factors, recent research provided theoretical and mechanistic groundings. Tutunov et al. (2023) proposed that CoT efficacy stems from the model's ability to approximate the true conditional distribution of reasoning, where structured exemplars help the model infer the task's latent logic and reduce generation ambiguity. Prabhakar et al. (2024) refined this view through a controlled case study, characterizing CoT as a probabilistic process heavily modulated by output *probability*, task *memorization* in training data, and step-wise *complexity*. Mechanistically, Wu et al. (2023) revealed how specific components of the CoT prompt drive model generation via gradient-based feature attribution.

Takeaway: CoT prompting yields significant gains primarily in tasks involving *mathematical, logical, or symbolic reasoning*. Its efficacy depends more on the structural coherence and relevance of exemplars, the length of reasoning, and the prompt phrasing. Several theoretical and mechanistic frameworks were proposed to understand such driving factors.

3.2 How Does Chain-of-Thought Remodel Internal Computation?

Chain-of-thought prompting does more than just alter an LLM's output format. Growing evidence shows that it fundamentally changes the model's internal computation into a "reasoning mode", where the model retrieves and updates information in a stepwise fashion, leveraging the intermediate computational steps as external memory.

The emergence of iteration heads. First, Cabannes et al. (2024) identified the "iteration head" — an attention head that emerges during CoT. These heads explicitly focus on the model's previously generated tokens to carry forward interim results. For example, in a loop counter task, an iteration head attends to the token "Step 4" to generate "Step 5". This effectively allows the model to create a virtual *recurrent neural network* (RNN) where the hidden state is externalized as text. In another study of a *Llama-2* model (Touvron et al., 2023) solving multi-step ontology queries, Dutta et al. (2024) also identified early-layer attention heads that "move information along ontological relationships" in the contexts that are relevant to the current sub-problem. The emergence of iteration heads provides supporting evidence that CoT enables the model to internally utilize generated text as an external memory for sequential reasoning.

Evidence of state maintenance and update. Besides the access to external memory, studies show that LLMs with CoT can also maintain and update dynamic internal states to track the reasoning process. Zhang et al. (2025a) found that when using CoT for state-tracking tasks, LLMs embed an implicit finite state automaton in their hidden layers. Specific feed-forward neurons in later layers were found to correspond directly to discrete problem states, forming a circuit that reliably updates with each new reasoning step. This internal state representation is highly robust and works correctly even with noisy or incomplete CoT steps, suggesting the model learns a resilient state-updating algorithm. By probing individual neurons of LLMs, Rai and Yao (2024) offered more granular evidence of state maintenance. They identified specific "reasoning neurons" in Llama-2's feed-forward layers that activate to hold partial results, such as carried values during arithmetic. Their activation helps explain why including particular steps (*e.g.*, an explicit breakdown of a sum) in the CoT prompt is effective: they reliably trigger the neurons responsible for maintaining the intermediate state.

Computational depth matters more than token semantics. Notably, the internal process of sequential reasoning appears to persist even when the CoT rationale lacks semantic meaning. For example, Pfau et al. (2024) replaced the meaningful CoT text with filler tokens (*e.g.*, "..."). Surprisingly, models could still solve complex reasoning tasks simply by generating these dots. Similarly, Goyal et al. (2024) found that introducing a learnable "pause" token significantly boosts performance on tasks from QA to math. These findings suggest that the semantic content of reasoning steps may be secondary to the computational time they buy. The sheer act of generating extra tokens (regardless of their meaning) provides necessary computational depth; each token grants the model an additional forward pass through all its layers. This extra "think time" enables the model to implement complex reasoning algorithms that cannot be executed in a single pass. Bharadwaj (2024) reinforced this interpretation through a mechanistic study. They found that even when CoT steps are replaced by placeholders, the model's deeper layers still encode the missing steps, which can be recovered to their correct semantic content via a logit lens probe.

Parallelism and reasoning shortcuts. Finally, although growing evidence reveals the sequential nature of CoT's internal computation, other studies have found that *LLMs often run multiple reasoning pathways in parallel during CoT*, meaning that the model's internal reasoning process is not strictly sequential. For example, Dutta et al. (2024) identified a "functional rift" where the model simultaneously tries to solve the problem directly from the question ("reasoning shortcuts") while also following the step-by-step procedure, and these parallel approaches then converge in later layers. Nikankin et al. (2025) found that models perform arithmetic via a "bag of heuristics" (many simple feature detectors) rather than a single step-by-step algorithm. Arcuschin et al. (2025) observed that the models can still arrive at the correct answer, even if they might make a mistake in an early step internally. The above evidence on parallelism and shortcuts reveals that CoT's internal workings are more complicated. It is a combination of sequential step-by-step reasoning, parallel associative shortcuts, and occasional after-the-fact rationalizations.

Takeaway: CoT activates a robust "reasoning mode" where models leverage generated tokens as external memory to execute stepwise internal computation, including fetching and carrying forward intermediate results and updating internal states. This core process persists even when CoT rationales are hidden or nonsensical. Yet, this internal computation is not strictly sequential but a parallel process involving multiple pathways and shortcuts.

3.3 Why CoT Enhances Reasoning Abilities?

Empirically, explicit reasoning with CoT often solves complex tasks more accurately than implicit latent reasoning. Several reasons have been identified for why CoT prompting dramatically improves reasoning performance.

CoT augments computational expressiveness. Recent theoretical studies demonstrate that CoT enhances transformers' expressiveness and computational capacity, enabling them to solve problems in higher complexity classes. A standard transformer decoder without CoT performs constant-depth computation per token, limiting it to the complexity class TC^0 (Merrill and Sabharwal, 2023a,b; Chiang et al., 2023). Such models theoretically cannot solve inherently serial problems because the required computation depth grows with input size, while the model's depth is fixed. CoT breaks this limit. By feeding the output back into the input, CoT allows the transformer to simulate an RNN or a Turing Machine. The effective depth of the computation becomes proportional to the length of the generated chain. This elevates the transformer's expressiveness to Polynomial Time (P) (Merrill and Sabharwal, 2024), making inherently serial or recursive computations solvable where they otherwise are not (Feng et al., 2023; Li et al., 2024b; Kim and Suzuki, 2025; Bavandpour et al., 2025).

CoT introduces modularity that reduces sample complexity. CoT decomposes complex tasks into granular, independent sub-problems. This modularity provides an inductive bias that matches the structure of complex, multi-step problems, enabling the model to master tasks with significantly less data. Through both experimental and theoretical evidence, Li et al. (2023) demonstrated that CoT decouples in-context learning into a "filtering" phase and a "learning" phase that significantly reduces the sample complexity required to learn compositional structures like MLPs. Extending this learnability perspective, Yang et al. (2025a) demonstrated that CoT can render inherently "unlearnable" tasks efficiently learnable by reducing the sample complexity of the overall task to that of its hardest individual reasoning step. Wen et al. (2025) further identified that this efficiency stems from the *sparse sequential dependencies* among tokens. CoT induces interpretable, sparse attention patterns that enable polynomial sample complexity, whereas implicit reasoning requires exponentially many samples to disentangle dense dependencies.

CoT enables more robust reasoning. First, evidence has been found that CoT *promotes robust generalization* by encouraging models to learn generalizable solution patterns rather than overfitting to surface-level statistical shortcuts. For example, Yao et al. (2025a) demonstrated that CoT-trained models induce a two-stage generalizing circuit that internalizes the reasoning process, leading to strong OOD generalization even in the presence of training noise. Complementing this, Li et al. (2025a) provided a theoretical guarantee for CoT generalization, showing that CoT maintains high performance even when context examples are noisy or erroneous, as it relies on step-by-step pattern matching rather than fragile input-output mappings. Second, CoT helps *reduce the propagation of errors* during reasoning. Gan et al. (2025) identified a "snowball error effect" in implicit reasoning, where minor inaccuracies accumulate into significant failures. They demonstrated that CoT-based strategies mitigate this by expanding the reasoning search space, which effectively lowers the probability of cumulative information loss and prevents errors from cascading through the reasoning chain.

Takeaway: CoT enhances LLM reasoning capabilities through three primary mechanisms: 1) It breaks the constant-depth limitation of the standard transformer, extending its effective computational depth. 2) It introduces modularity and inductive bias more aligned with multi-step reasoning, thus reducing the sample complexity required to learn complex tasks. 3) It facilitates robust generalization to OOD data and mitigates error propagation.

3.4 Does Chain-of-Thought Equate to Explainability?

Explicit reasoning appears to provide transparency, leading users to assume that CoT explanations accurately reveal how the model arrived at an answer. However, substantial evidence indicates that CoT outputs often do not faithfully reflect the model's actual decision-making process (Lanham et al., 2023; Turpin et al., 2023; Chen et al., 2025c; Barez et al., 2025), a phenomenon referred to as the *unfaithfulness* of CoT reasoning.

Evidence of CoT unfaithfulness. Recent studies reveal that CoT frequently functions as *post-hoc rationalization* rather than the causal driver of predictions (Kudo et al., 2024; Arcuschin et al., 2025; Lewis-Lim et al., 2025). For instance, Turpin et al. (2023) demonstrated that models often alter their predictions based on spurious cues, such as the reordering of multiple-choice options. In such cases, the models still tend to confabulate logical-sounding CoT rationales that hide the actual spurious cause of their decision. Similarly, when correct answers are injected as hints, models often invent spurious derivations to support the injected answer without acknowledging the hint's influence (Chen et al., 2025c). Furthermore, mechanistic analyses uncovered "silent error corrections", where models internally correct mistakes without updating the CoT rationale (Arcuschin et al., 2025). Unfaithfulness is also evident in *sycophancy*, where models prioritize agreement with user beliefs over truthfulness. Even when models possess the correct internal knowledge, they frequently concede to incorrect user premises and generate plausible rationales to justify these compliant responses (Sharma et al., 2024). Collectively, these findings highlight a fundamental disconnect between verbalized rationales and internal computations, challenging the premise that CoT equates to explainability.

Mechanistic understanding of CoT unfaithfulness. Recent mechanistic analyses attribute this unfaithfulness to a fundamental mismatch between the distributed, parallel nature of transformer computation and the sequential nature of explicit reasoning. As discussed in Section 3.2, many works have revealed the *distributed nature* of LLMs' internal reasoning; transformer-based LLMs frequently employ multiple redundant computational pathways to process information, *e.g.*, simultaneously leveraging memorization, heuristics, and algorithmic circuits (McGrath et al., 2023; Dutta et al., 2024; Nikankin et al., 2025). Consequently, CoT only acts as a "lossy projection" of high-dimensional internal states, often capturing only a fraction of the model's actual decision process (Dutta et al., 2024). Because computation is highly distributed, a single CoT rationale can capture at most one of many simultaneous causal pathways. As a result, CoTs typically omit influential factors and serve only as partial, post-hoc rationalisations of the model's underlying distributed, superposed computation (Barez et al., 2025). This architectural dissonance makes unfaithfulness difficult to mitigate. Tanneru et al. (2024) demonstrated that even when training objectives explicitly penalize inconsistency, models still revert to plausible-but-not-causal explanations on complex tasks, highlighting the inherent difficulty in eliciting faithful CoT reasoning from LLMs.

Takeaway: While chain-of-thought offers the appearance of transparency, it does not equate to faithful explainability. CoT often functions as post-hoc rationalization rather than a true reflection of the model's internal processing. Mechanistically, this unfaithfulness stems from a structural mismatch between the distributed, parallel computation of transformers and the sequential nature of explicit reasoning.

4 Future Research Directions

Rigorous causal analysis in real-world settings. A fundamental challenge in current mechanistic research is *the disparity between idealized experimental settings and the complexities of real-world reasoning*. First, the reliance on toy models and synthetic data limits the generalizability of current findings. For example, while the "grokking" phenomenon has been identified as a potential pathway for the emergence of implicit multi-hop reasoning, most empirical evidence is derived from toy models trained from scratch on synthetic tasks (§ 2.2). Consequently, it remains an open question whether the phase transitions observed in these controlled environments truly govern the development of reasoning capabilities in foundation models trained on large-scale, naturalistic corpora.

Second, the field should move beyond correlational analysis, which only proves information presence, to *rigorous causal verification* within these complex settings. Unlike clean synthetic environments, real-world data is ubiquitous with spurious cues, making it difficult to distinguish genuine reasoning circuits from robust shortcut heuristics (§ 2.3). Therefore, causal interventions are crucial for proving that identified internal representations are truly drivers of correct inference in the wild. This understanding should ideally translate into *robust training-time interventions* that penalize such shortcuts, forcing models to learn generalizable algorithms despite the noisy data distribution. Ultimately, future work must aim to synthesize these insights into a unified theoretical framework that explains how diverse components, from memorization circuits to reasoning heads, interact within the massive scale of foundation models.

Bridging the faithfulness gap of explicit CoT reasoning. As discussed in § 3.4, a critical bottleneck in current LLMs is the "functional rift" (Dutta et al., 2024) between the model's internal, parallel processing and its sequential, explicit CoT reasoning. This structural mismatch forces models to compress high-dimensional, distributed latent states into a low-bandwidth stream of discrete tokens, often resulting in CoT that functions as a post-hoc rationalization rather than a causal driver. To address this, future research must explore *white-box alignment methods* that enforce a causal link between implicit and explicit reasoning. Promising avenues include developing training objectives that penalize discrepancies between the model's hidden states (its true decision process) and its generated rationale (Wang et al., 2025a,c), imposing architectural constraints that compel the model to rely solely on the generated CoT for subsequent steps (Viteri et al., 2024), as well as "self-explaining" dense internal representations into faithful natural language steps (Sengupta and Rekik, 2025). Further exploration of these directions is critical for aligning explicit outputs with internal dynamics, ensuring CoT serves as a valid window into the model's computation.

Mechanistic understanding of Latent CoT reasoning. Beyond the dichotomy of implicit and explicit CoT, an emerging paradigm is *latent CoT reasoning* (Chen et al., 2025b; Li et al., 2025b), where models are designed to simulate explicit reasoning trajectories entirely within hidden states. Unlike standard implicit reasoning, which relies on the fixed depth of a standard transformer, latent CoT architectures often introduce additional computational capacity via continuous "thought tokens", iterative refinement, or recurrent state updates, frequently learning these behaviors by distilling explicit CoT data into latent representations. This approach theoretically offers the best of both worlds: it broadens the model's expressive capacity and computational depth while eliminating the redundant decoding costs of natural language tokens.

While various latent CoT architectures have been proposed (Hao et al., 2024; Mitra et al., 2024; Shen et al., 2025), mechanistic interpretability has lagged significantly behind these innovations. While a vast body of work has explored the latent reasoning mechanisms of *standard transformers* (§ 2.1), research into the internal dynamics of these novel latent CoT models remains limited (Zhang and Viteri, 2024; Wang et al., 2025b; Zhang et al., 2025b). Critical open questions remain: Does distilling explicit CoT truly force the model to internalize a sequential, step-by-step reasoning process, or does the model collapse the teacher's rationale into high-dimensional statistical shortcuts? Therefore, gaining more mechanistic insights is crucial for designing next-generation latent CoT architectures and training objectives that effectively combine the interpretability of explicit reasoning with the efficiency of implicit computation.

White-box evaluation metrics for LLM reasoning. As we gain a deeper mechanistic understanding of multi-step reasoning, it should guide the development of evaluation protocols that go beyond simple end-task accuracy. Current black-box metrics (*e.g.*, final accuracy) are increasingly insufficient, as models frequently arrive at correct answers via non-robust shortcuts, statistical heuristics, or "bag-of-words" processing (§ 2.3). To rigorously distinguish genuine reasoning from sophisticated pattern matching, the field requires *"white-box" evaluation metrics* that integrate model internals into the evaluation protocol. Pioneering efforts have begun to explore this direction. For example, Cao et al. (2025) introduced a mechanism-interpretable metric (MUI) that quantifies the "effort" required to solve a task, defined as the proportion of activated neurons or features. A truly capable model should achieve higher performance with lower effort. While this area remains under-explored, developing metrics that not only score the final output but also verify the presence of necessary internal computational signatures, such as the formation of bridge entities in intermediate layers (Yang et al., 2025c), is a crucial future trend. By defining reasoning not just as the correct outcome but as the execution of a verified internal process, we can prevent the overestimation of model capabilities and ensure that improvements on leaderboards reflect true algorithmic generalization.

From mechanistic interpretation to model control. While current research has successfully identified various reasoning circuits, such as iteration heads or deduction heads, most work remains observational. A major frontier for future study is the shift towards *pragmatic interpretability* (Nanda et al., 2025a,b), moving from passively explaining mechanisms to actively leveraging them for model control and editing, a paradigm closely aligned with *Representation Engineering (RepE)* (Wehner et al., 2025). For example, if we can reliably identify the specific components responsible for multi-step logic, *e.g.*, the state-maintenance neurons identified by (Rai and Yao, 2024), we can potentially intervene in real-time to correct reasoning errors or suppress shortcut neurons (Ju et al., 2024). Such interventions enable the development of "self-correcting" architectures that actively monitor internal states to detect and resolve failures like "silent errors" on the fly. Ultimately, this enables a transition from interpretability as a passive analysis tool to an active, foundational component for robust and safe reasoning systems.

5 Conclusion

In this survey, we provided a comprehensive overview of the mechanisms underlying multi-step reasoning in large language models. We structured our analysis around two fundamentally distinct computational paradigms: implicit reasoning and explicit reasoning. Through a framework of seven interconnected research questions, we systematically explored the internal dynamics of latent inference, the emergence of reasoning capabilities, and the mechanistic impact of chain-of-thought prompting on model computation and expressiveness. Despite significant progress in opening the black box, critical challenges remain. Looking ahead, we outlined a roadmap for future research, emphasizing the necessary shift from passive observation to causal intervention and the need for rigorous verification in real-world settings to build more reliable reasoning systems.

Limitations

This survey concentrates strictly on the mechanistic understanding of multi-step reasoning within transformer-based LLMs. Consequently, we do not cover other aspects of reasoning, such as probabilistic inference, creative planning, or commonsense reasoning, which may operate under different mechanistic principles. Additionally, our scope is limited to the current paradigm of text-based transformers; we do not extensively address reasoning mechanisms in Multimodal LLMs (MLLMs), alternative architectures like Diffusion Language Models (DLMs), or neural networks that predate the modern era of large language models.

References

Roman Abramov, Felix Steinbauer, and Gjergji Kasneci. 2025. Grokking in the wild: Data augmentation for real-world multi-hop reasoning with transformers. In *Proceedings of the 42th International Conference on Machine Learning (ICML)*. OpenReview.net.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-thought reasoning in the wild is not always faithful. *CoRR*, abs/2503.08679.

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. 2025. Chain-of-thought is not explainability. *CoRR*.

Alireza Amiri Bavandpour, Xinting Huang, Mark Rofin, and Michael Hahn. 2025. Lower bounds for chain-of-thought reasoning in hard-attention transformers. In *Proceedings of the 42th International Conference on Machine Learning (ICML)*. OpenReview.net.

Aryasomayajula Ram Bharadwaj. 2024. Understanding hidden computations in chain-of-thought reasoning. *CoRR*, abs/2412.04537.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 14113–14130. Association for Computational Linguistics.

Jannik Brinkmann, Abhay Sheshadri, Victor Levoso, Paul Swoboda, and Christian Bartelt. 2024a. A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4082–4102.

Jannik Brinkmann, Abhay Sheshadri, Victor Levoso, Paul Swoboda, and Christian Bartelt. 2024b. A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4082–4102. Association for Computational Linguistics.

Vivien Cabannes, Charles Arnal, Wassim Bouaziz, Xingyu Yang, François Charton, and Julia Kempe. 2024. Iteration head: A mechanistic study of chain-of-thought. In *Proceedings of the 2024 Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Yixin Cao, Jiahao Ying, Yaoning Wang, Xipeng Qiu, Xuanjing Huang, and Yugang Jiang. 2025. Model utility law: Evaluating llms beyond performance through mechanism interpretable metric. *CoRR*, abs/2504.07440.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025a. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *CoRR*, abs/2503.09567.

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. 2025b. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *CoRR*, abs/2505.16782.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vladimir Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025c. Reasoning models don't always say what they think. *CoRR*, abs/2505.05410.

David Chiang, Peter Cholak, and Anand Pillay. 2023. Tighter bounds on the expressivity of transformer encoders. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 5544–5562.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024a. Navigate through enigmatic labyrinth A survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1173–1203.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024b. Navigate through enigmatic labyrinth A survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1173–1203.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. 2024. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *Transactions on Machine Learning Research (TMLR)*, 2024.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Proceedings of the 2023 Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Zeyu Gan, Yun Liao, and Yong Liu. 2025. Rethinking external slow-thinking: From snowball errors to probability of correct reasoning. In *Proceedings of the 42th International Conference on Machine Learning (ICML)*. OpenReview.net.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the*

41th International Conference on Machine Learning (ICML)*. OpenReview.net.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training language models with pause tokens. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. Deepseek-coder: When the large language model meets programming - the rise of code intelligence. *CoRR*, abs/2401.14196.

Tianyu Guo, Hanlin Zhu, Ruiqi Zhang, Jiantao Jiao, Song Mei, Michael I. Jordan, and Stuart Russell. 2025. How do llms perform two-hop reasoning in context? *CoRR*, abs/2502.13913.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *CoRR*, abs/2412.06769.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023a. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4902–4919.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023b. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4902–4919. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1830–1842. Association for Computational Linguistics.

Tianjie Ju, Yijin Chen, Xinwei Yuan, Zhuosheng Zhang, Wei Du, Yubin Zheng, and Gongshen Liu. 2024. Investigating multi-hop factual shortcuts in knowledge editing of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8987–9001. Association for Computational Linguistics.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of co-occurrence on factual knowledge of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7721–7735. Association for Computational Linguistics.

Juno Kim and Taiji Suzuki. 2025. Transformers provably solve parity efficiently with chain of thought. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Keito Kudo, Yoichi Aoki, Tatsuki Kuribayashi, Shusaku Sone, Masaya Taniguchi, Ana Brassard, Keisuke Sakaguchi, and Kentaro Inui. 2024. Think-to-talk or talk-to-think? when llms come up with an answer in multi-step reasoning. *CoRR*, abs/2412.01113.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702.

Samuel Lewis-Lim, Xingwei Tan, Zhixue Zhao, and Nikolaos Aletras. 2025. Analysing chain of thought dynamics: Active guidance or unfaithful post-hoc rationalisation? *CoRR*, abs/2508.19827.

Hongkang Li, Songtao Lu, Pin-Yu Chen, Xiaodong Cui, and Meng Wang. 2025a. Training nonlinear transformers for chain-of-thought inference: A theoretical generalization analysis. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. 2025b. Implicit reasoning in large language models: A comprehensive survey. *CoRR*, abs/2509.02350.

Yingcong Li, Kartik Sreenivasan, Angeliki Giannou, Dimitris Papailiopoulos, and Samet Oymak. 2023. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. In *Proceedings of the 2023 Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Zhaoyi Li, Gangwei Jiang, Hong Xie, Linqi Song, Defu Lian, and Ying Wei. 2024a. Understanding and patching compositional reasoning in llms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9668–9688. Association for Computational Linguistics.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024b. Chain of thought empowers transformers to solve inherently serial problems. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Ziyue Li, Chenrui Fan, and Tianyi Zhou. 2025c. Where to find grokking in LLM pretraining? monitor memorization-to-generalization without test. *CoRR*, abs/2506.21551.

Tianhe Lin, Jian Xie, Siyu Yuan, and Deqing Yang. 2025. Implicit reasoning in transformers is reasoning through shortcuts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9470–9487. Association for Computational Linguistics.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. 2024. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *CoRR*, abs/2410.21333.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What makes chain-of-thought prompting effective? A counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535. Association for Computational Linguistics.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. 2023. The hydra effect: Emergent self-repair in language model computations. *CoRR*, abs/2307.15771.

William Merrill and Ashish Sabharwal. 2023a. A logic for expressing log-precision transformers. In *Proceedings of the 2023 Annual Conference on Neural Information Processing Systems (NeurIPS)*.

William Merrill and Ashish Sabharwal. 2023b. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545.

William Merrill and Ashish Sabharwal. 2024. The expressive power of transformers with chain of thought. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, pages 14420–14431. IEEE.

Neel Nanda, Josh Engels, Arthur Conmy, Senthooran Rajamanoharan, Bilal Chughtai, Callum McDougall, János Kramár, and Lewis Smith. 2025a. A pragmatic vision for interpretability. AI Alignment Forum.

Neel Nanda, Josh Engels, Senthooran Rajamanoharan, Arthur Conmy, Bilal Chughtai, Callum McDougall, János Kramár, and Lewis Smith. 2025b. How can interpretability researchers help AGI go well? LessWrong.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2025. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

nostalgebraist. 2020. interpreting GPT: the logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`. LessWrong.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *CoRR*, abs/2209.11895.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Jacob Pfau, William Merrill, and Samuel R. Bowman. 2024. Let's think dot by dot: Hidden computation in transformer language models. *CoRR*, abs/2404.15758.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *CoRR*, abs/2201.02177.

Akshara Prabhakar, Thomas L. Griffiths, and R. Thomas McCoy. 2024. Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3710–3724. Association for Computational Linguistics.

Daking Rai and Ziyu Yao. 2024. An investigation of neuron activation as a unified lens to explain chain-of-thought eliciting arithmetic reasoning of llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7174–7193.

Nikta Gohari Sadr, Sangmitra Madhusudan, and Ali Emami. 2025. Think or step-by-step? unzipping the black box in zero-shot prompts. *CoRR*, abs/2502.03418.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. 2025. Reasoning with latent thoughts: On the power of looped transformers. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Prajit Sengupta and Islem Rekik. 2025. X-node: Self-explanation is all we need. In *Reconstruction and Imaging Motion Estimation, and Graphs in Biomedical Image Analysis - First International Workshop, RIME 2025, and 7th International Workshop, GRAIL 2025*, volume 16150 of *Lecture Notes in Computer Science*, pages 184–194. Springer.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. CODI: compressing chain-of-thought into continuous space via self-distillation. *CoRR*, abs/2502.21074.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research (TMLR)*, 2023.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2024. Chain of thoughtlessness? an analysis of cot in planning. In *Proceedings of the 2024 Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. 2024. On the hardness of faithful chain-of-thought reasoning in large language models. *CoRR*, abs/2406.10625.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Rasul Tutunov, Antoine Grosnit, Juliusz Ziomek, Jun Wang, and Haitham Bou-Ammar. 2023. Why can large language models generate correct chain-of-thoughts? *CoRR*, abs/2310.13571.

Scott Viteri, Armand Lamparth, Pierre Chatain, and Clark Barrett. 2024. Markovian agents for informative language modeling. *CoRR*, abs/2404.18988.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2717–2739.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. In *Proceedings of the 2024 Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Xiangqi Wang, Yue Huang, Yujun Zhou, Xiaonan Luo, Kehan Guo, and Xiangliang Zhang. 2025a. Causally-enhanced reinforcement policy optimization. *CoRR*, abs/2509.23095.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F. Wong, and Rui Wang. 2025b. Latent space chain-of-embedding enables output-free LLM self-evaluation. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Zijian Wang, Yanxiang Ma, and Chang Xu. 2025c. Eliciting chain-of-thought in base llms via gradient-based representation optimization. *Preprint*, arXiv:2511.19131.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. 2025. Taxonomy, opportunities, and challenges of representation engineering for large language models. *Transactions on Machine Learning Research (TMLR)*, 2025.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. 2025. From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. 2023. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. *CoRR*, abs/2307.13339.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024a. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Chenxiao Yang, Zhiyuan Li, and David Wipf. 2025a. Chain-of-thought provably enables learning the (otherwise) unlearnable. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024b. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10210–10229. Association for Computational Linguistics.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. 2025b. Do large language models perform latent multi-hop reasoning without exploiting shortcuts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3971–3992.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. 2025c. Do large language models perform latent multi-hop reasoning without exploiting shortcuts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3971–3992. Association for Computational Linguistics.

Zhipeng Yang, Junzhuo Li, Siyu Xia, and Xuming Hu. 2025d. Internal chain-of-thought: Empirical evidence for layer-wise subtask scheduling in llms. *CoRR*, abs/2505.14530.

Xinhao Yao, Ruifeng Ren, Yun Liao, and Yong Liu. 2025a. Unveiling the mechanisms of explicit cot training: How chain-of-thought enhances reasoning generalization. *CoRR*, abs/2502.04667.

Yuekun Yao, Yupei Du, Dawei Zhu, Michael Hahn, and Alexander Koller. 2025b. Language models can learn implicit multi-hop reasoning, but only if they have lots of training data. *CoRR*, abs/2505.17923.

Jiaran Ye, Zijun Yao, Zhidian Huang, Liangming Pan, Jinxin Liu, Yushi Bai, Amy Xin, Liu Weichuan, Xiaoyin Che, Lei Hou, and Juanzi Li. 2025. How does transformer learn implicit reasoning? In *Proceedings of the 2025 Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*, San Diego, USA.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Yijiong Yu. 2025. Do llms really think step-by-step in implicit reasoning? *CoRR*, abs/2411.15862.

Zeping Yu, Yonatan Belinkov, and Sophia Ananiadou. 2025. Back attention: Understanding and enhancing multi-hop reasoning in large language models. *CoRR*, abs/2502.10835.

Jason Zhang and Scott Viteri. 2024. Uncovering latent chain of thought vectors in language models. *CoRR*, abs/2409.14026.

Yifan Zhang, Wenyu Du, Dongming Jin, Jie Fu, and Zhi Jin. 2025a. Finite state automata inside transformers with chain-of-thought: A mechanistic study on state tracking. In *Proceedings of the 63nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 13603–13621. Association for Computational Linguistics.

Yuyi Zhang, Boyu Tang, Tianjie Ju, Sufeng Duan, and Gongshen Liu. 2025b. Do latent tokens think? a causal and adversarial analysis of chain-of-continuous-thought. *CoRR*, abs/2512.21711.

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, and Zhi-Qin John Xu. 2025c. Complexity control facilitates reasoning-based compositional generalization in transformers. *CoRR*, abs/2501.08537.