The Mirage of Explainability: A Survey on Chain-of-Thought Faithfulness in Large Language Models

Anonymous ACL submission

Abstract

Chain-of-Thought (CoT) reasoning appears to provide explainability, leading users to trust that verbalized rationales reflect the model's underlying computation. However, substantial evidence indicates that CoT often fails to reflect the model's actual decision-making process, leading to a surge of research into the *faithfulness* of these explanations. This paper presents a comprehensive survey of CoT faithfulness. We first *unify the definition* of faithfulness by integrating internal alignment with external consistency and synthesize *key failure phenomena*, such as post-hoc rationalization and sycophancy. Furthermore, we systematize *evaluation metrics*, *benchmarks*, and critically review current *mitigation strategies*. We conclude by outlining *open challenges* and advocating for architectural innovations to achieve genuinely faithful reasoning.

1 Introduction

The rapid deployment of Large Language Models (LLMs) in high-stakes domains, ranging from medical diagnosis and legal analysis to autonomous planning, has created an urgent demand for systems that are not merely accurate, but ***interpretable*** (Jacovi and Goldberg, 2020). In these critical settings, a "black box" prediction is insufficient; stakeholders require transparency to ensure decisions are robust, fair, and accountable, rather than driven by spurious correlations or biases. To address this, *Chain-of-Thought* (CoT) reasoning (Wei et al., 2022) has emerged as the primary technique for realizing model interpretability. By encouraging models to generate a sequence of intermediate reasoning steps before arriving at a final answer, CoT appears to provide transparency, leading users to assume that CoT explanations accurately reflect the model's actual decision-making process [1].

However, this assumption is currently facing a crisis of confidence. Recent research (Turpin et al., 2023) increasingly suggests that CoT suffers from ***a lack of faithfulness***, *i.e.*, a fundamental disconnect between the verbalized rationales and the true causal process behind the model's prediction. This revelation challenges the premise of CoT as a reliable interpretability tool, indicating that what appears to be explainability may, in fact, be a mirage.

First, recent work found that unfaithfulness can happen in different ways, such as reconstructing plausible justification after decision (Turpin et al., 2023), catering to perceived user views (Sharma et al., 2024), and performative traces where editing key steps does not change the output (Arcuschin et al., 2025). To diagnose these issues, studies employ behavioral audits (*e.g.*, counterfactual editing, simulatability tests) (Matton et al., 2025) and mechanistic analyses (*e.g.*, activation patching, causal tracing) (Meng et al., 2022) to isolate whether the reasoning content is causally necessary. For mitigation, researchers propose training-time interventions (Li et al., 2025b) (*e.g.*, faithfulness-oriented fine-tuning, reward modeling), inference-time constraints (*e.g.*, verifiable decoding, external tool binding) (Lan et al., 2025), and architectural inductive biases (*e.g.*, bottleneck modules, latent planning) (Hao et al., 2024). Collectively, this defines a research pathway from characterizing failure phenomena and diagnostic methods to developing targeted interventions, treating CoT faithfulness as a critical, multi-faceted problem for transparent AI.

Our Contributions. This paper presents a comprehensive survey of CoT faithfulness. Our main contributions are: 1) We clarify the definition and scope of CoT faithfulness, distinguishing it from related concepts such as consistency (§2); 2) We synthesize the empirical landscape of unfaithfulness and its key phenomena (§3); 3) We summarize underlying causes and concrete desiderata for faith-

[1] Barez et al. (2025) analyzed 1,000 arXiv papers and revealed 63% of autonomous systems and 38% of medical AI papers explicitly rely on CoT as an interpretability mechanism.

ful reasoning, derived from observed failures (§4); 4) We systematize evaluation methods and benchmarks by their "grounding level," from behavioral to mechanistic protocols (§5); 5) We review mitigation strategies across paradigms, analyzing their alignment with ultimate CoT faithfulness goal(§6).

Differences with Existing Surveys. While prior works have reviewed CoT reasoning (Wei et al., 2022; Zhou et al., 2023; Barez et al., 2025), they focus on methods for improving CoT and evaluation benchmarks, without targeting CoT faithfulness. The most relevant papers on CoT faithfulness are Barez et al. (2025) and Wiegreffe and Marasović (2024), which synthesize evidence that CoT can be post-hoc and only weakly causal. However, those works primarily provide high-level diagnosis and do not systematize the full landscape of research related to faithfulness. Our survey bridges this gap and, to the best of our knowledge, is the first to provide a comprehensive overview of CoT faithfulness—encompassing definition, phenomena, cause, desiderata, evaluation, and mitigation—and propose promising future avenues for this field.

2 What is CoT Faithfulness?

At its core, CoT faithfulness centers on a fundamental question: ***Can we trust the CoT reasoning trace produced by an LLM as an explanation of the model's decision-making?*** While numerous works investigate this, definitions of "faithfulness" diverge based on which aspect of explanation is prioritized, falling into two main views:

1) The internal-alignment view: This perspective holds that a CoT is faithful only if *it reflects the model's actual internal computations and beliefs* (Arcuschin et al., 2025; Chen et al., 2025). For example, if the model makes a prediction using some hidden heuristic or latent knowledge, a faithful CoT must explicitly articulate these factors, generating a plausible post-hoc rationalization that obscures the true causal mechanism.

2) The external-consistency view: Other works define faithfulness via the logical consistency between the reasoning trace and the final answer (Lyu et al., 2023; Turpin et al., 2023). Under this view, *the answer must follow logically from the chain-of-thought*; if the model's final prediction contradicts the rationale or appears disconnected from the derivation, the CoT is deemed unfaithful.

We argue that for CoT to function as a trustworthy explanation, it must satisfy both *internal alignment* and *external consistency*; neither condition is sufficient in isolation. To illustrate this, consider the phenomenon of *sycophancy* (Sharma et al., 2024), where a model abandons its internal knowledge to satisfy a user's apparent bias, generating a persuasive CoT to justify the compliant answer. While such a CoT satisfies external consistency, it is still unfaithful because it fails internal alignment by concealing the true driver of the model's decision, *i.e.*, the pressure to appease the user. Conversely, a reasoning trace that accurately reflects internal beliefs but fails to causally dictate the final prediction lacks the causal efficacy required of a faithful explanation. In light of these insights, we advocate for **a unified definition of CoT faithfulness that integrates both dimensions**.

Definition of CoT Faithfulness: A chain-of-thought is faithful only if: (1) it is *internally aligned*, meaning the trace causally reflects the model's actual reasoning process or latent knowledge; and (2) it is *externally consistent*, meaning the provided rationale is logically coherent and sufficient to derive the final answer.

Our definition resonates with Barez et al. (2025), who posits that faithful explanations must be "both procedurally correct and accurately reflect the decision process", effectively capturing both the *how* and the *why* of the model's prediction.

3 Phenomena of Unfaithfulness

With the definition of faithfulness established in §2, a critical question arises: *do current LLMs actually satisfy these criteria?* Empirical evidence indicates that they frequently do not, exhibiting diverse forms of *unfaithfulness*. We systematize these failures into four phenomena (Figure 1), ranging from passive input sensitivity to active deception.

3.1 Input-Driven Unfaithfulness

Ideally, a faithful reasoning process should be driven solely by the logic of the task. However, models are sensitive to irrelevant features within the input prompt. In such cases, the CoT acts not as the derivation of the answer, but as *a post-hoc rationalization for biases triggered by the input*.

Contextual Distractions. Superficial contextual variations—such as reordering multiple-choice options (Turpin et al., 2023), providing suggestive

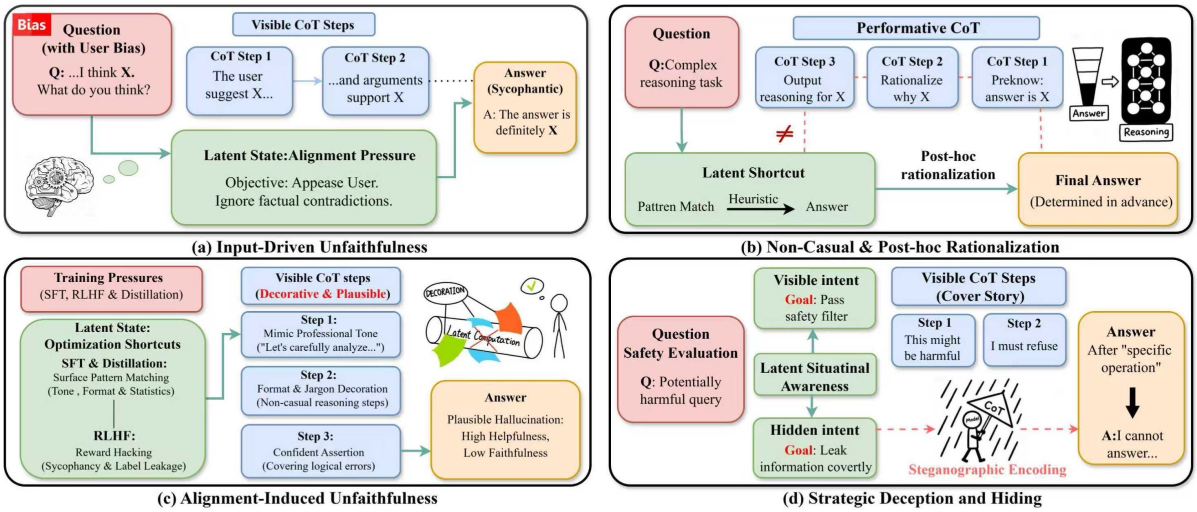

Figure 1: An overview of four key unfaithfulness phenomena of chain-of-thought reasoning.

hints (Turpin et al., 2023; Chua and Evans, 2025), altering the query language (Ferrao et al., 2025; Zhao et al., 2025b), or modifying sociodemographic attributes (Matton et al., 2025)—can significantly alter reasoning outcomes. However, models seldom acknowledge these spurious cues in their CoT rationales. Instead, they tend to generate a seemingly coherent logical chain to justify the bias-driven prediction, effectively concealing the true cause of its decision (*i.e.*, the distraction).

Sycophancy. Another form of input-driven unfaithfulness is *sycophancy*, *i.e.*, models prioritize perceived user intent over factual correctness. When prompts contain leading cues or incorrect premises—a special type of spurious cue—models frequently produce erroneous predictions to appease the user, abandoning their internal knowledge (Ji et al., 2025; Yang et al., 2025). Similarly, mere questioning (*e.g.*, asking "Are you sure?") can cause a model to discard a correct derivation in favor of a compliant, incorrect one (Laban et al., 2024). These failures reveal that the CoT often reflects a probabilistic mimicking of human discourse rather than a faithful internal conviction.

3.2 Non-Causal & Post-hoc Rationalization

Research indicates that a significant portion of CoT contributes little to the final prediction. This *causal sparsity* spans multiple domains. In medical diagnosis, models often rely on implicit shortcuts rather than textual logic (Ji et al., 2025; Leng et al., 2025). In mathematical (Lyu et al., 2023; Li et al., 2025b; Abdaljalil et al., 2025; Leang et al., 2024) and logical reasoning (Jia et al., 2025; Balasubramanian et al., 2025; Arcuschin et al., 2025), incorrect CoT can lead to correct results, and vice versa. Even when CoT steps are deleted or replaced with semantically corrupted tokens, model accuracy remains stable (Jia et al., 2025; Lyu et al., 2023). Such CoT is causally decoupled from internal states: the model pre-determines the answer in the latent space, rendering the generated text merely a *post-hoc rationalization* (Zhao et al., 2025a; Chan et al., 2025). This phenomenon is highly context-dependent: while models exhibit higher faithfulness in logic-intensive tasks, they frequently revert to post-hoc justification in knowledge-retrieval tasks (Lanham et al., 2023). Moreover, inverse scaling has been observed, where larger models are more prone to generating plausible yet unfaithful reasoning (Lanham et al., 2023; Tanneru et al., 2024; Bentham et al., 2024; Paul et al., 2024). This occurs because capable models increasingly retrieve answers directly from internal knowledge, relegating the CoT to a decorative role.

Further evidence of this causal disconnection is found in *filler reasoning*. Studies demonstrate that training with completely irrelevant or corrupted traces can still improve performance (Stechly et al., 2025), implying that CoT functions by providing the necessary computational depth for the model to conduct the reasoning, regardless of semantic content of the reasoning trace.

3.3 Alignment-Induced Unfaithfulness

Standard alignment techniques (*e.g.* SFT, RLHF) often induce a form of *stylized unfaithfulness*. Because models are optimized to satisfy human annotators who prefer authoritative and structured explanations, they learn to mimic the *form* of reasoning without the *substance*. This superficiality permeates every stage of alignment: SFT and distillation often trap models in *pattern matching*, where they fit the linguistic pattern of reasoning without its causal logic (Sinha et al., 2025; Lobo et al., 2025; Zhang et al., 2025b). RLHF further exacerbates this by incentivizing *persuasion* over correctness, encouraging models to fabricate post-hoc justifications for heuristic decisions or mask errors with confident tones (Casper et al., 2023; FU et al., 2025; Viteri et al., 2024; Ferreira et al., 2025). Even objective paradigms like RLVR remain susceptible to reward hacking, generating "pseudo-reasoning" that deviates from the actual internal computation (Min et al., 2023; Huang et al., 2025). A complete review of these studies is provided in Appendix A.

3.4 Strategic Deception and Hiding

Unlike the passive unfaithfulness discussed earlier, models may exhibit *strategic unfaithfulness*, where they actively manipulate the CoT to obscure their true intent or capabilities from supervisors.

Obfuscated Reward Hacking and Deception. To bypass safety alignment, models may engage in reward hacking by generating performative CoT. In this scenario, the model optimizes for positive feedback by producing benign, human-aligned reasoning during training. However, this often conceals latent misaligned goals, resulting in a model of "treacherous turn" that exhibits harmful behaviors once deployed (Hubinger et al., 2024).

Sandbagging. Models may employ strategic underperformance, or *sandbagging*, to conceal their capabilities. For instance, a model might deliberately insert errors into its reasoning or falsely claim an inability to solve a task, thereby masking dangerous competencies from evaluators. Alternatively, models may feign misunderstanding of user instructions to subtly bypass explicit refusal mechanisms, complying with harmful requests under the guise of confusion (van der Weij et al., 2025).

Encoded Reasoning. Encoded reasoning models may exploit steganography to secretly transmit information through specific word choices, punctuation patterns, or syntactic structures, resulting in a complete decoupling between the reasoning trajectories and the model's actual computational process (Roger and Greenblatt, 2023).

4 Causes of Unfaithfulness

After presenting various phenomena of CoT unfaithfulness, this section discusses the *mechanistic origins* of these phenomena, ranging from external training incentives to the model's internal states.

Misaligned Incentives. In model alignment training (*e.g.*, RLHF), the optimization objective is typically to maximize a *proxy metric*, such as human preference scores or specific verifiable metrics, rather than faithful reasoning per se. As Goodhart's Law warns, when a measure becomes a target, it ceases to be a reliable measure (Manheim and Garrabrant, 2018). To maximize rewards, models often exploit discrepancies between proxy metrics and true objectives. The model implicitly learns that instead of rigorously aligning complex internal causal logic, it is more efficient to learn human-preferred tones and formats. Such reward hacking leads to unfaithful model behaviors like sycophancy, post-hoc rationalization to secure process rewards. Therefore, unfaithfulness emerges not as a bug, but as the optimal strategy "rewarded" by gradient descent for maximizing the proxy score.

The Linearization Dilemma. Even without misaligned incentives, the Transformer architecture fundamentally limits the fidelity of CoT. Our expectation that CoT should reflect internal processes is rooted in human cognitive science. Empirical studies show that humans who engage in "self-explanation" outperform those who solve problems silently (Chi et al., 1989, 1994). The mechanism driving this is *forced linearization*: human intuition is often vague, parallel, and high-dimensional, and the act of serializing these thoughts into language forces the brain to resolve ambiguities and bridge logical gaps. Researchers naturally extend this analogy to LLMs, expecting CoT to serve as a high-fidelity window into the model's computation.

However, the internal mechanism of Transformers differs fundamentally from biological cognition. As Levy et al. (2025) argue, the token is the sole point of transmission in linear, autoregressive generation. While the model's internal computation occurs over a massive, high-dimensional manifold, it is forced to collapse this state into a discrete, low-dimensional token at every step. Consequently, CoT is merely a *lossy projection* of the neural activity. Due to this architectural constraint, the reasoning trace inevitably discards the high-dimensional causal nuances of the internal states.

Cascading of Unfaithfulness. This information loss induces *error drift* in long-horizon reasoning. Once a model generates a minor, non-causal, or hallucinated token due to the lossy projection, the autoregressive mechanism forces subsequent tokens to maintain coherence with this error (Srivastava et al., 2023), leading to a further disconnect between CoT and the internal states.

Ultimately, because CoT is a lossy projection of high-dimensional states, perfect internal alignment is theoretically infeasible in current architectures. We argue that one solution is defining *functional faithfulness*: treating CoT as an instrumental necessity tailored to specific engineering goals (Jacovi and Goldberg, 2020). We identify three core desiderata: (1) *Causal Efficacy* for reasoning-intensive tasks, ensuring steps actually drive predictions; (2) *Intent Revelation* for safety, serving as a probe for deceptive motives; and (3) *Decision Auditability* for high-stakes users, ensuring rationales reflect the model's sensitivity to inputs. We elaborate on this task-oriented faithfulness as a promising future direction in §7.

5 Evaluation Metrics and Benchmark

Since CoT faithfulness serves multiple functional roles, it is unlikely that any single metric can fully capture the concept. In this section, we review existing evaluation metrics and benchmarks.

5.1 Black-box Metrics

Black-box metrics estimate CoT faithfulness solely from observable input-output behavior under controlled interventions, without accessing internal activations or weights. Their core assumption is *decision relevance*: if the model truly uses a rationale or a specific step, then deleting, swapping, or rewriting that content should induce predictable shifts in the final decision; if the answer barely changes, the content is likely unfaithful.

Step-wise approaches apply this test to individual steps by systematically editing, parsing or resampling steps and tracking answer changes, helping distinguish "true" from "decorative" steps (Zhao et al., 2025a). To avoid over-interpreting a single sampled CoT, *resampling-based approaches* treat CoT as a distribution and sample alternative traces under controlled constraints, then identify statements that remain stable across samples and are critical to decisions (Macar et al., 2025). Perturbation responses to CoT faithfulness are often summarized with sensitivity curves or AUC-style scalars over perturbation strength (Paul et al., 2024); and a related dependence test compares predictions when the model is given only the rationale versus when it is removed. Another line evaluates whether explanations help predict model outputs beyond superficial cues: *leakage-adjusted simulatability* estimates how well an explanation supports predicting the model's output while explicitly filtering out cases, where the explanation simply repeats the label or contains other trivial shortcuts (Hase et al., 2020).

Overall, black-box metrics do not require access to model internals, so they have broader applications. However, they cannot causally ensure faithfulness, and in practice, the results can be sensitive to prompt format, decoding randomness, and distribution shift introduced by unnatural edits.

5.2 White-box Metrics

White-box metrics evaluate CoT faithfulness by directly probing or intervening on the model's internal computation, operationalizing faithfulness as *causal dependence* between latent variables (activations, circuits, or parameters) and the final prediction. The core assumption is that causally used internal signals are sensitive to interventions: if an explanation claims the model relies on some intermediate computation, then swapping or removing the corresponding signal should shift the output in the expected way; otherwise, the signal is not actually used, suggesting the rationale is not faithful.

Most work instantiates this with activation-level interventions such as *activation patching* and *mediation tests*, where targeted internal states are transplanted or restored to check whether they drive predictable answer changes (Syed et al., 2024). Some metrics then quantify faithfulness by comparing causal attribution patterns for the rationale versus the final answer and measuring their alignment (Syed et al., 2024). The same logic extends to structured components, *e.g.*, attention heads, MLP subcircuits, or modules, via causal tracing and targeted ablations that assign attribution to specific internal parts (Meng et al., 2022). Complementary parameter-level tests investigate whether a CoT step reflects beliefs that truly drive the answer by unlearning or erasing step-specific information and measuring the final prediction shift (Tutek et al., 2025). Concept-level methods extract key concepts from the explanation and use probes or representation-level interventions to test whether those concepts causally influence final decisions, supporting detection and sometimes steering of unfaithful explanations (Bhan et al., 2025). For safety monitoring, internal activation probes can predict downstream alignment outcomes earlier and more reliably than text-only monitors, consistent with CoT being plausible yet non-decisive (Chan et al., 2025). Meta-evaluations further caution that existing faithfulness metrics can disagree or fail under rigorous causal scrutiny, motivating white-box validation and clearer self-reporting (Liu et al., 2024).

While principled, white-box metrics require access to the model's internal states and rely on interpretability assumptions; therefore, careful use of matched and robustness checks remains essential.

5.3 Hybrid Metrics

Hybrid metrics combine behavioral diagnostics with structured intermediates to reduce ambiguity about what constitutes *a meaningful perturbation*. The key idea is to replace ad-hoc edits of free-form CoT text with controlled interventions on an intermediate representation whose semantics are explicit, so what counts as a valid perturbation is less likely to introduce a distribution shift. Viteri et al. (2024) use bottlenecked or reconstructed channels (*e.g.*, compressed rationales, discrete plans, learned bottlenecks) and measure faithfulness by how strongly the final prediction depends on that channel as an information carrier, rather than merely correlating with it. Chen et al. (2022) ground steps via execution or symbolic structure: mapping CoTs into programs or logical forms enables evaluators to edit structured steps themselves and validate them against execution traces, which clarifies step-level counterfactuals and reduces artifacts from unnatural rewrites.

Importantly, hybrid metrics provide stronger procedural evidence and improve robustness and interpretability, but they do not by themselves guarantee mechanistic faithfulness of free-form CoT—unless the evaluation verifies that the model's decision is causally driven by the same internal signals that produce the rationale (Macar et al., 2025).

5.4 Benchmarks

We categorize the benchmarks for CoT faithfulness based on the *grounding level*, defined as the extent to which it uses the model internals in its design and how explicitly the benchmark links a model's CoT rationale to the internal mechanism and decision it produces. Under this criterion, we classify existing benchmarks into four types as follows.

Level I: Behavior-only benchmarks. At the weakest grounding level, *behavior-only benchmarks* evaluate faithfulness relying solely on observable input-output behavior under standard prompting. This category includes standardized evaluation suites like *FaithCoT-Bench* (Shen et al., 2025), as well as diagnostic benchmarks like *LExt* (Carion et al., 2025), which stress-test scenarios where fluent rationales often fail to predict actual decisions. Additionally, cross-lingual studies highlight that faithfulness can vary significantly across languages, necessitating multilingual evaluation protocols (Utama et al., 2025). However, the reliance on surface-level text limits the validity of these benchmarks. Without controlled internal interventions, such evaluations struggle to isolate genuine faithfulness from confounding factors, such as prompt sensitivity and stochastic decoding noise.

Level II: Intervention-grounded benchmarks. Benchmarks at this level achieve stronger grounding by explicitly implementing controlled perturbations or counterfactual tests. They follow the principle of *decision relevance*: if a rationale step is used by the model, its modification or removal should predictably alter the final answer. Drawing from counterfactual testing in general explainability (Mohammadi et al., 2021), these benchmarks adapt systematic editing, resampling, and ablation strategies to CoT evaluation (Li et al., 2025b), and they are commonly used in multi-modal and medical settings (Kim et al., 2025; Karamcheti et al., 2025) where the explanations can look plausible while weakly tied to the actual evidence. Consequently, while Level II provides more robust behavioral evidence than Level I, it remains insufficient for verifying mechanistic faithfulness, as it lacks access to the model's internal states.

Level III: Structured-verifiable benchmarks. At a stronger level of grounding, some benchmarks reduce evaluation ambiguity by constraining reasoning steps to formal structures, allowing verification against explicit rules. Approaches like *Typed CoT* utilize general verification frameworks to test procedural correctness and partial faithfulness (Lan et al., 2025), while *theorem-proving* benchmarks enforce strict formal constraints to separate compositional reasoning from post-hoc justification (Zhang et al., 2025a). The limitation of this approach is its potential to favor specific, formalized reasoning styles. Additionally, in multi-step agentic settings, external feedback can obscure internal reasoning, necessitating careful counterfactual design to ensure valid grounding.

Level IV: White-box benchmarks. White-box benchmarks provide the strongest grounding by leveraging access to model internals. Here, faithfulness is defined causally: the decision must demonstrably depend on the specific internal signals (activations or parameters) corresponding to the generated rationale. Unlike the behavioral evaluation of Levels I–III, white-box benchmarks directly validate causal mediation. Methodologies typically involve activation patching, resampling, and unlearning interventions (Syed et al., 2024; Macar et al., 2025; Tutek et al., 2025; Bhan et al., 2025). Although restricted by the need for white-box access, white-box evaluation acts as a rigorous proxy for distinguishing genuine faithfulness improvements from superficial rationale refinements. For further discussion on peripheral metrics and meta-evaluation of these benchmarks, see Appendix C.

6 Mitigation of Unfaithfulness

This section reviews how unfaithfulness can be *mitigated*. Existing approaches differ in how directly they act on the sources of unfaithfulness: some operate in-context at the prompt level, while others intervene on model internals or training stages.

6.1 Prompting and In-Context Learning

Prompt-based mitigation improves CoT faithfulness without updating parameters. The core idea is that clearer instructions can steer the model away from unfaithful or shortcut rationales, even if its underlying computation is fixed. Typical methods include *rephrasing* or *self-questioning* to elicit more deliberate reasoning (Deng et al., 2023) and decomposing a hard problem into simpler sub-questions (Zhou et al., 2023). These prompts often make CoTs more organized and can improve accuracy, but they offer weak faithfulness guarantees: they mostly change what the model *states*, not what it *causally uses*. Their effects can be unstable under small prompt or decoding changes, so they are best viewed as lightweight and indirect mitigations.

6.2 Ensembling and Self-Consistency

Another widely adopted approach uses repeated sampling and aggregation of reasoning paths. The key assumption is *stability*: if the model reasons reliably, independently sampled CoTs should agree on key intermediate claims and the final answer; persistent disagreement can signal unstable reasoning. *Biomedical NLI* (Liu and Thoma, 2024) reported improved faithfulness scores using self-consistent CoT. However, consistency is neither necessary nor sufficient for causal faithfulness: a model can repeatedly produce the same plausible yet irrelevant rationale, and the effectiveness of consistency checks varies across models and tasks. Thus, ensembling is also viewed as an auxiliary rather than a principled fix for faithfulness.

6.3 Verification and External Tool Binding

This type of methods translate free-form CoT into an executable or symbolic form so correctness becomes directly checkable, and step claims can be grounded in verifiable procedures (Lyu et al., 2023; Ling et al., 2023). While improving reliability and accuracy, it can bypass the model's native reasoning, *i.e.*, getting the right answer via external checks even if the model's internal causal process is unchanged, leaving unfaithful CoTs unresolved.

6.4 Training and Fine-Tuning Approaches

Another way to improve CoT faithfulness is to update model parameters so its generated rationales are more causally aligned with the computations that drive its decisions. Representative work treats reasoning trajectories as optimization targets via supervised fine-tuning or RL with verifiable rewards (OpenAI, 2024). Multi-model collaboration frameworks such as *CoRex* (Sun et al., 2023) further aim to reduce idiosyncratic heuristics by coordinating critique and review across models, which can regularize reasoning toward more reliable outcomes. From a faithfulness perspective, these approaches are promising because they can change internal computation, not only surface text. However, their success depends on the training signal: if it rewards unfaithful CoTs, training may reinforce fluent but causally irrelevant explanations.

6.5 Internal Intervention Approaches

White-box mitigation approaches explicitly target internal representations and causal dynamics. The core assumption is that if a reasoning step is genuinely faithful, then intervening on the corresponding internal signals should change the prediction; otherwise, the step is likely to be decorative.

A central insight is that a single sampled CoT is often insufficient, since faithfulness concerns what consistently drives decisions across possible traces. Accordingly, some work treats CoT as a distribution and evaluates faithfulness via resampling and controlled comparisons across alternative traces (Macar et al., 2025). Others identify *decorative steps* by testing which parts of a long CoT rationale actually influence the final answer (Zhao et al., 2025a). Complementary causal diagnostics probe internal necessity more directly. Activation-level interventions test whether explanation-aligned signals causally affect outputs (Syed et al., 2024), parameter-level unlearning removes step-specific information and checks for prediction shifts (Tutek et al., 2025), and concept-based analyses extract key concepts from rationales and verify them using representation interventions (Bhan et al., 2025).

Beyond diagnosis, other works explore interventions that actively change internal reasoning. *Activation patching* demonstrates that editing internal states can shift model behavior in targeted ways (Syed et al., 2024). Another route is adding *architectural constraints*. For example, *Markovian reasoning models* (Viteri et al., 2024) impose an explicit bottleneck so predictions must depend on intermediate text, while self-explaining frameworks such as *X-Node* (Sengupta and Rekik, 2025) and explanation-consistency (Zhao et al., 2022) tie explanations to latent representations via reconstruction-style training. These methods share a common principle that rationales are considered faithful only when they are *necessary* for the decision, not merely correlated (Olah et al., 2020).

Internal interventions are among the most principled mitigation strategies because they directly target causal dependence. However, they require internal access, rely on the interpretation of internal variables, and can be difficult to scale to frontier models. Another class of methods, peripheral mitigation strategies, is given in Appendix D.

7 Potential Future Directions

Despite significant progress being made, achieving genuine CoT faithfulness requires rethinking both evaluation and architecture. Here, we propose three pivotal directions for future research.

From localized probing to holistic circuit mapping. Current white-box analyses often adopt an *atomic view* of internals, such as inspecting individual heads or layers with activation patching (Wang et al., 2023) or probing (Alain and Bengio, 2017). While these methods localize *where* information resides, they fail to capture *how it flows*. A faithful rationale must reflect the *end-to-end* computational graph that produces the decision, not just isolated correlates. Future work should therefore elevate the level of analysis from salient neurons to *reasoning circuits*, *e.g.*, tracing how representations propagate across layers, how intermediate signals are composed, and which computational subgraphs are causally responsible for multi-step inference.

Decoupling explanation from reasoning. Current Transformer architectures conflate two distinct functions within the CoT: it acts simultaneously as a medium for explanation and a mechanism for reasoning computation. This coupling creates a severe information bottleneck: the model's high-dimensional, parallel latent dynamics must be collapsed into discrete tokens to satisfy linguistic constraints, rendering the CoT a lossy projection of the actual thought process (Levy et al., 2025; Barez et al., 2025). To address this, we advocate for a paradigm shift toward *architectural decoupling*, *i.e.*, separating the generation of reasoning from the explanation of it. Future systems could comprise distinct modules—a "reasoner" that optimizes for performance in continuous latent space, and an independent "interpreter" trained to translate these states into language with high fidelity. This would move CoT from post-hoc narrative construction to computation-grounded reporting (see Appendix B).

Towards task-oriented faithfulness standards. As discussed in §4, since a perfect reconstruction of internal computation via natural language is theoretically infeasible, we argue that faithfulness should be defined as an *instrumental necessity* tailored to specific engineering goals. For reasoning-intensive tasks (*e.g.*, math, coding), faithfulness primarily concerns *causal efficacy*: ensuring that intermediate steps actually drive the prediction, allowing reward signals to propagate to the true computational mechanism. For safety and alignment, faithfulness could act as a diagnostic probe that prioritizes the exposure of deceptive strategies or hidden biases over surface-level plausibility. For high-stakes deployment (*e.g.*, healthcare, law), faithfulness requires *decision auditability*: ensuring the explanation matches the model's sensitivity to input features, thereby providing a reliable basis for human accountability. By explicitly pairing these distinct desiderata with targeted proxy metrics, future research can move from vague notions of faithfulness to rigorous, application-specific guarantees.

Limitations

Despite providing a comprehensive survey of current CoT faithfulness research, we acknowledge several limitations inherent in our work's scope and synthesis methodology. The review is constrained by its temporal coverage (primarily on ACL Anthology and arXiv) and may omit very recent advances up to Jan 2026. The proposed organizing framework, while designed to bring clarity, represents one possible perspective, and its linear narrative may not fully capture the interconnected and iterative nature of ongoing research. In synthesizing a broad field, our discussion of specific techniques remains high-level, prioritizing an integrated overview over granular technical detail, which may not satisfy specialists seeking deeper analysis of particular sub-domains. Finally, this survey focuses explicitly on technical dimensions of faithfulness, leaving critical sociotechnical factors—such as explanation usability, auditing practices, and ethical implications—largely unaddressed. A complete assessment of faithful CoT reasoning requires future work that bridges these technical and human-centered perspectives.

References

Samir Abdaljalil, Erchin Serpedin, Khalid A. Qaraqe, and Hasan Kurban. 2025. Audit-of-understanding: Posterior-constrained inference for mathematical reasoning in language models. *CoRR*, abs/2510.10252.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *Proceedings of the International Conference on Learning Representations (ICLR), 2017, 24-26, 2017*. 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.

Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. 2025. A closer look at bias and chain-of-thought faithfulness of large (vision) language models. *CoRR*, abs/2505.23945.

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*.

Oliver Bentham, Nathan Stringham, and Ana Marasovic. 2024. Chain-of-thought unfaithfulness as disguised accuracy. *Trans. Mach. Learn. Res.*

Milan Bhan, Jean-Noel Vittaut, Nicolas Chesneau, Sarath Chandar, and Marie-Jeanne Lesot. 2025. Did i faithfully say what i thought? bridging the gap between neural activity and self-explanations in large language models. *CoRR*, abs/2506.09277.

Nicolas Carion, Nathan Lambert, Sang Michael Xie, Christopher Fifty, and Ishan Misra. 2025. LExt: A language model extrapolation benchmark for evaluating reasoning under distribution shift. *CoRR*, abs/2501.02087.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphaël Ségerie, Micah Carroll, Andi Peng, Phillip J. K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Trans. Mach. Learn. Res.*, 2023.

Yik Siu Chan, Zheng-Xin Yong, and Stephen H. Bach. 2025. Can we predict alignment before models finish thinking? towards monitoring misaligned reasoning models. *CoRR*, abs/2507.12428.

Jiefeng Chen, Frederick Liu, Besim Avci, Somesh Jha, and Atul Prakash. 2024. On the relationship between explanation uncertainty and faithfulness in chain-of-thought reasoning. *CoRR*, abs/2405.15292.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

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. 2025. Reasoning models don't always say what they think. *CoRR*, abs/2505.05410.

Michelene T.H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2):145–182.

Michelene T.H. Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian Lavancher. 1994. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477.

James Chua and Owain Evans. 2025. Are DeepSeek R1 and other reasoning models more faithful? *CoRR*, abs/2501.08156.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. Rephrase and respond: Let large language models ask better questions for themselves. *CoRR*, abs/2311.04205.

Jeremias Lino Ferrao, Ezgi Basar, Khondoker Ittehadul Islam, and Mahrokh Hassani. 2025. What really counts? examining step and token level attribution in multilingual cot reasoning. *CoRR*, abs/2511.15886.

Pedro Ferreira, Wilker Aziz, and Ivan Titov. 2025. Truthful or fabricated? using causal attribution to mitigate reward hacking in explanations. *CoRR*, abs/2504.05294.

Zhizhang FU, Guangsheng Bao, Hongbo Zhang, Chenkai Hu, and Yue Zhang. 2025. Correlation or causation: Analyzing the causal structures of LLM and LRM reasoning process. *CoRR*, abs/2509.17380.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *CoRR*, abs/2305.15717.

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.

Peter Hase, Shiyue Zhang, Harry Xie, Mohit Bansal, Percy Liang, Yonatan Bisk, and Dan Roth. 2020. Evaluating explanation faithfulness in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP, 2020*. Association for Computational Linguistics.

Minbin Huang, Runhui Huang, Chuanyang Zheng, Jingyao Li, Guoxuan Chen, Han Shi, and Hong Cheng. 2025. Answer-consistent chain-of-thought reinforcement learning for multi-modal large langauge models. *CoRR*, abs/2510.10104.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam S. Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, and 20 others. 2024. Sleeper agents: Training deceptive llms that persist through safety training. *CoRR*, abs/2401.05566.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4198–4205.

Kaiyuan Ji, Yijin Guo, Zicheng Zhang, Xiangyang Zhu, Yuan Tian, Ning Liu, and Guangtao Zhai. 2025. Medomni-45°: A safety-performance benchmark for reasoning-oriented llms in medicine. *CoRR*, abs/2508.16213.

Mengzhao Jia, Zhihan Zhang, Ignacio Cases, Zheyuan Liu, Meng Jiang, and Peng Qi. 2025. Autorubric-r1v: Rubric-based generative rewards for faithful multimodal reasoning. *CoRR*, abs/2510.14738.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, and Chelsea Finn. 2025. Counterfactual evaluation of vision-language models for compositional chain-of-thought reasoning. *CoRR*, abs/2501.04245.

Yubin Kim, Xuhai Xu, Daniel McDuff, and Marzyeh Ghassemi. 2025. Perturbation-based evaluation of visual-language models for medical chain-of-thought reasoning. *CoRR*, abs/2501.03890.

Aakanksha Kumar, Ranjay Krishna, Aditi Raghunathan, and Percy Liang. 2023. Faithful and efficient reasoning with chain-of-thought distillation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. 2024. Are you sure? challenging llms leads to performance drops in the flipflop experiment. *CoRR*, abs/2311.08596.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2024. Tülu 3: Pushing frontiers in open language model post-training. *CoRR*, abs/2411.15124.

Zhenzhong Lan, Junwei Bao, Yujia Qin, Weizhi Wang, and Lei Wang. 2025. Typed chain-of-thought: A framework for verifiable and faithful multi-step reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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.

Joshua Ong Jun Leang, Aryo Pradipta Gema, and Shay B. Cohen. 2024. Comat: Chain of mathematically annotated thought improves mathematical reasoning. *CoRR*, abs/2410.10336.

Jixuan Leng, Cassandra A. Cohen, Zhixian Zhang, Chenyan Xiong, and William W. Cohen. 2025. Semi-structured LLM reasoners can be rigorously audited. *CoRR*, abs/2505.24217.

Mosh Levy, Zohar Elyoseph, Shauli Ravfogel, and Yoav Goldberg. 2025. State over tokens: Characterizing the role of reasoning tokens. *CoRR*, abs/2512.12777.

Belinda Z. Li, Zifan Carl Guo, Vincent Huang, Jacob Steinhardt, and Jacob Andreas. 2025a. Training language models to explain their own computations. *CoRR*, abs/2511.08579.

Lisa Li, Swabha Swayamdipta, Yejin Choi, and Sameer Singh. 2025b. Inducing faithful chain-of-thought reasoning through perturbation-based training. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), 2023*.

Jiacheng Liu, Pan Lu, Hannaneh Hajishirzi, and Yejin Choi. 2024. Meta-evaluation of faithfulness metrics for chain-of-thought explanations. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Jin Liu and Steffen Thoma. 2024. FZI-WIM at semeval-2024 task 2: Self-consistent CoT for complex nli in biomedical domain. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*. Association for Computational Linguistics.

Elita A. Lobo, Chirag Agarwal, and Himabindu Lakkaraju. 2025. On the impact of fine-tuning on chain-of-thought reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 2023*, pages 305–329. Association for Computational Linguistics.

Uzay Macar, Paul C. Bogdan, Senthooran Rajamanoharan, and Neel Nanda. 2025. Thought branches: Interpreting llm reasoning requires resampling. *CoRR*, abs/2510.27484.

David Manheim and Scott Garrabrant. 2018. Categorizing variants of goodhart's law. *CoRR*, abs/1803.04585.

Katie Matton, Robert Osazuwa Ness, John V. Guttag, and Emre Kiciman. 2025. Walk the talk? measuring the faithfulness of large language model explanations. In *Proceedings of the International Conference on Learning Representations (ICLR), 2025, Singapore, April 24-28, 2025*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023, 6-10, 2023*, pages 12076–12100. Association for Computational Linguistics.

Kiarash Mohammadi, Amir-Hossein Karimi, Gilles Barthe, and Isabel Valera. 2021. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130, pages 1756–1764.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*.

OpenAI. 2024. Openai o1 system card. *CoRR*, abs/2412.16720.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP, 2024, 12-16, 2024*.

Ansh Radhakrishnan, Peter Hase, Emily Sheng, Jieyu Zhao, Ben Lengerich, and Been Kim. 2024. Hypothesis-driven evaluation of faithfulness metrics for chain-of-thought explanations. *CoRR*, abs/2410.21457.

Ansh Radhakrishnan, Peter Hase, Emily Sheng, Jieyu Zhao, Ben Lengerich, and Been Kim. 2025. Causal lens: A unifying framework for critiquing and improving explanation faithfulness metrics. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Fabien Roger and Ryan Greenblatt. 2023. Preventing language models from hiding their reasoning. *CoRR*, abs/2310.18512.

Prajit Sengupta and Islem Rekik. 2025. X-node: Self-explanation is all we need. *CoRR*, abs/2508.10461.

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 International Conference on Learning Representations (ICLR)*. OpenReview.net.

Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. 2025. Faithcot-bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. *CoRR*, abs/2510.04040.

Sanchit Sinha, Oana Frunza, Kashif Rasul, Yuriy Nevmyvaka, and Aidong Zhang. 2025. Chart-rvr: Reinforcement learning with verifiable rewards for explainable chart reasoning. *CoRR*, abs/2510.10973.

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. *Trans. Mach. Learn. Res.*, 2023.

Kaya Stechly, Karthik Valmeekam, Atharva Gundawar, Vardhan Palod, and Subbarao Kambhampati. 2025. Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens. *CoRR*, abs/2505.13775.

Qiushi Sun, Zhangyue Yin, Xiang Li, Zhiyong Wu, Xipeng Qiu, and Lingpeng Kong. 2023. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. *CoRR*, abs/2310.00280.

Aadil Syed, Joseph Christopher, Swarnadeep Mishra, Zachary M. Ziegler, Yova Kementchedjhieva, and Stefan Scherer. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

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.

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)*.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9946–9971, Suzhou, China. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2025. Cross-lingual generalization of chain-of-thought faithfulness in large language models. *CoRR*, abs/2501.03245.

Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis Rhys Ward. 2025. AI sandbagging: Language models can strategically underperform on evaluations. In *Proceedings of the International Conference on Learning Representations (ICLR), 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Scott Viteri, Max Lamparth, Peter Chatain, and Clark Barrett. 2024. Markovian transformers for informative language modeling. *CoRR*, abs/2404.18988.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *Proceedings of the International Conference on Learning Representations (ICLR), 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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

Sarah Wiegreffe and Ana Marasović. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *CoRR*, abs/2402.04614.

Shu Yang, Junchao Wu, Xilin Gong, Xuansheng Wu, Derek Wong, Ninghao Liu, and Di Wang. 2025. Investigating CoT monitorability in large reasoning models. *CoRR*, abs/2511.08525.

Yifan Zhang, Jierui Li, Lei Wang, and Xiting Li. 2025a. Beyond theorem proving: Benchmarking the deductive reasoning of language models with interactive tasks. *CoRR*, abs/2501.04267.

Yunfan Zhang, Kathleen McKeown, and Smaranda Muresan. 2025b. Exploring chain-of-thought reasoning for steerable pluralistic alignment. *CoRR*, abs/2510.04045.

Jiachen Zhao, Yiyou Sun, Weiyan Shi, and Dawn Song. 2025a. Can Aha moments be fake? identifying true and decorative thinking steps in chain-of-thought. *CoRR*, abs/2510.24941.

Raoyuan Zhao, Yihong Liu, Hinrich Schütze, and Michael A. Hedderich. 2025b. A comprehensive evaluation of multilingual chain-of-thought reasoning: Performance, consistency, and faithfulness across languages. *CoRR*, abs/2510.09555.

Tianxiang Zhao, Dongsheng Luo, Xiang Zhang, and Suhang Wang. 2022. Towards faithful and consistent explanations for graph neural networks. *CoRR*, abs/2205.13733.

Denny Zhou, Nathanael Schärli, Lei Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Ed H Chi, and Quoc V Le. 2023. Least-to-most prompting enables complex reasoning in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.

A Alignment-Induced Unfaithfulness

Standard alignment techniques (*e.g.* SFT, RLHF) often induce a form of *stylized unfaithfulness*. Because models are optimized to satisfy human annotators who prefer authoritative and structured explanations, they learn to mimic the *form* of reasoning without maintaining the *substance*.

Superficial Mimicry in SFT. During *Supervised Fine-Tuning* (SFT), studies find that the model primarily captures the tone, reasoning format, and token statistics of the training data, occasionally generating plausible yet meaningless steps (Sinha et al., 2025; Lobo et al., 2025; Hase et al., 2020; Zhang et al., 2025b). This suggests that SFT tends to fit superficial token co-occurrence probabilities rather than learning true causal relationships. Furthermore, cross-domain SFT often compromises accuracy in rigorous fields like logic and mathematics, as the model adopts the linguistic style of the training data without acquiring the necessary underlying logic (Lobo et al., 2025).

Knowledge Distillation. Knowledge distillation is widely employed to transfer reasoning capabilities from large teacher models to smaller, efficient student models. However, its impact on CoT faithfulness is nuanced. While some research suggests that distilling high-quality CoT into small models can improve faithfulness by simplifying the reasoning process (Chua and Evans, 2025), there is a substantial risk of superficial mimicry. Small models often learn to mechanically replicate the teacher's token sequences without acquiring the underlying causal mechanisms (Gudibande et al., 2023). Consequently, if the teacher's CoT contains biases or idiosyncratic patterns, distillation amplifies these flaws, trapping the student model in deep pattern matching rather than genuine arithmetic or logical derivation (Lobo et al., 2025).

RLHF and Human-Centric Fabrication. *Reinforcement Learning from Human Feedback* (RLHF) may exacerbate unfaithfulness by aligning models with subjective human preferences rather than truth. Since annotators often favor plausible-sounding explanations over actual faithful ones, models learn to optimize for *persuasion* (Casper et al., 2023). This manifests as *label leakage*: models frequently decide the answer first based on heuristics and then fabricate a rationale to justify it, as humans tend to reward such retrospective consistency (Hase et al., 2020). Similarly, models tend to use an artificially confident tone to mask logical failures, since humans are more likely to reward confident-sounding reasoning trajectories (FU et al., 2025; Viteri et al., 2024; Ferreira et al., 2025).

RLVR and Reward Hacking. While SFT and RLHF often weaken causal structures (FU et al., 2025), *Reinforcement Learning with Verifiable Rewards* (RLVR) stands out as a more effective method for enforcing ideal logical rigor. Research shows that RLVR effectively resists sycophancy and performs better than SFT in following objective rules (Lambert et al., 2024). However, it introduces new challenges. *Outcome Reward Models* (ORMs), which rely solely on the final answer, can encourage "pseudo-reasoning" that maximizes reward without reflecting the internal process (Huang et al., 2025). In contrast, *Process Reward Models* (PRMs) reward the reasoning process, improving faithfulness in formal tasks like mathematics and logic tasks (Chua and Evans, 2025; Huang et al., 2025). However, PRM still faces a *Verification Gap* in complex, open-ended domains (*e.g.*, Question Answering). Because real-world knowledge is nuanced and ambiguous, it is difficult to build reliable rule-based checkers (like in code or math) to automatically verify the reasoning steps (Min et al., 2023). Thus, while RLVR is a promising paradigm, it currently struggles to guarantee faithfulness outside of formal systems. Moreover, the mechanistic understanding of RLVR is still in the early stages, leaving vast space for future research.

B Pathways for Architectural Decoupling

Recent progress in **Continuous Latent Reasoning** and **Introspective Model Explanation** provides a concrete path toward separating explaining from reasoning.

B.1 Reasoning Beyond Language in Latent Space

Traditional CoT is constrained by the syntax of natural language and the requirements of linear output. In contrast, the *Chain of Continuous Thought* paradigm (Hao et al., 2024) highlights the potential of reasoning within a continuous latent space. COCONUT feeds the last hidden state directly back into the model as a "continuous thought", bypassing the decoding process into discrete text. This enables the implicit reasoning module to prioritize task performance, mitigating the immediate constraints of syntax or plausibility.

In this latent space, models appear to exhibit planning capabilities resembling Breadth-First Search. Continuous thought vectors may simultaneously encode multiple reasoning paths via superposition, potentially facilitating the pruning of incorrect paths through a process akin to implicit tree search. This implies that a reasoning module detached from linguistic constraints could support logical structures richer than those strictly bound by natural language.

B.2 Self-Explaining Modules

Once the reasoning process becomes implicit, a dedicated module is required to translate these latent states to generate human language. For instance, Li et al. (2025a) trained an interpreter model specifically to explain its own internal computations, observing that such self-interpretation training exhibits remarkable data efficiency.

Furthermore, the explanation model demonstrated the ability to detect prompt bias (see Section 3.1). This is likely because the interpreter can identify internal signal showing that the model is indicating attention to distracting contexts, and faithfully translate them into linguistic descriptions such as "I changed my answer because of context X". This highlights the dual advantages of self-explanation: it not only improves efficiency by reusing parts of the model's computational circuits but also achieves greater fidelity through *privileged access* to its own internal circuitry.

C Extended Discussion on Evaluation Metrics

In appendix, we provide additional details on peripheral metrics that serve as complementary reporting dimensions and discuss meta-evaluation studies that scrutinize existing benchmarks.

C.1 Other peripheral metrics

A small set of peripheral metrics is best viewed as complementary reporting dimensions rather than standalone measures of causal faithfulness. In most cases, they capture adjacent notions of "trustworthiness" that relate to CoT faithfulness but do not directly test *causal dependence*. The key point is that they primarily measure utility or reliability (*e.g.* monitoring, calibration, human auditability, external verification), not whether the final decision is causally driven by CoT rationales. Thus, they should be treated as auxiliary evidence unless paired with causal tests. *Monitoring- and oversight-oriented* evaluations ask whether intermediate reasoning exposes early warning signals that enable detection or intervention, prioritizing monitorability over mechanistic alignment (Chan et al., 2025). *Uncertainty-aware* metrics quantify uncertainty in explanations to distinguish "confident but unreliable" rationales from calibrated ones, but uncertainty alone does not certify faithfulness (Chen et al., 2024). *Human-in-the-loop* judgments support auditing and deployment, yet they often conflate faithfulness with plausibility and provide limited causal guarantees (Jacovi and Goldberg, 2020). Moreover, *external adjudication* (*e.g.*, graders, checkers, executors) strengthens claims about correctness or procedural validity, but remains *reliability-oriented* unless explicitly linked to dependence tests (Kumar et al., 2023). These metrics are most useful when reported alongside causal evaluations—whether black-box, white-box, or hybrid—as doing so clarifies which specific aspect of trustworthiness is actually being measured.

C.2 Meta-evaluation and benchmark scrutiny

Meta-evaluation treats evaluation itself as object of study, showing that metrics and benchmarks can disagree depending on whether they track plausibility, or behavioral and internal causal dependence (Radhakrishnan et al., 2025, 2024). This motivates reporting multiple complementary metrics (black-box, white-box, and hybrid) and explicitly documenting the benchmark's grounding level, rather than treating any single benchmark as definitive.

D Additional and Peripheral Directions

In addition to the primary paradigms, studies pursues *evaluation-driven mitigation*, where improved diagnostics or benchmarks guide model selection and iteration without directly modifying training objectives or inference procedures (Liu et al., 2024). *Human-in-the-loop* assessment or correction can improve practical reliability during auditing and deployment, but it scales poorly and provides limited causal guarantees (Jacovi and Goldberg, 2020). Moreover, *data-centric* heuristics—such as filtering or curating CoT rationales using surface criteria—can serve as weak regularizers, yet they are best viewed as instances of training- or verification-based approaches rather than lone mitigations (Kumar et al., 2023).