Chain-of-Thought Is Not Explainability

Fazl Barez[*] Oxford, WhiteBox
Tung-Yu Wu WhiteBox
Iván Arcuschin Independent
Michael Lan Independent
Vincent Wang Oxford, Cosmos
Noah Siegel Google DeepMind, UCL
Nicolas Collignon Independent
Clement Neo NTU
Isabelle Lee USC
Alasdair Paren Oxford
Adel Bibi Oxford
Robert Trager Oxford
Damiano Fornasiere Mila
John Yan WhiteBox
Yanai Elazar AI2, UW
Yoshua Bengio Mila

Abstract

Chains-of-thought (CoT) allow language models to verbalise multi-step rationales before producing their final answer. While this technique often boosts task performance and offers an impression of transparency into the model's reasoning, we argue that rationales generated by current CoT techniques can be misleading and are *neither necessary nor sufficient* for trustworthy interpretability. By analysing *faithfulness* in terms of whether CoTs are not only human-interpretable, but also reflect underlying model reasoning in a way that supports responsible use, we synthesise evidence from previous studies. We show that verbalised chains are frequently *unfaithful*, diverging from the true hidden computations that drive a model's predictions, and giving an incorrect picture of how models arrive at conclusions. Despite this, CoT is increasingly relied upon in high-stakes domains such as medicine, law, and autonomous systems—our analysis of 1,000 recent CoT-centric papers finds that ~25% explicitly treat CoT as an interpretability technique—and among them, papers in high-stakes domains specifically hinge on such interpretability claim heavily. Building on prior work in interpretability, we make three proposals: (i) avoid treating CoT as being sufficient for interpretability without additional verification, while continuing to use CoT for its communicative benefits, (ii) adopt rigorous methods that assess faithfulness for downstream decision-making, and (iii) develop causal validation methods (e.g., activation patching, counterfactual interventions, verifier models) to ground explanations in model internals.

1 Introduction

Chains-of-thought (CoT) enables large language models (LLMs) to generate a sequence of intermediate reasoning steps before arriving at a final answer [78, 44, 75]. It emerges either explicitly through direct instructions to "think step-by-step", or implicitly through reward signals or examples that demonstrate step-by-step reasoning. This technique has contributed to remarkable problem-solving abilities such as mathematics [48] and commonsense reasoning [71].

[*]Corresponding author: fazl@robots.ox.ac.uk

Preprint. Under review.

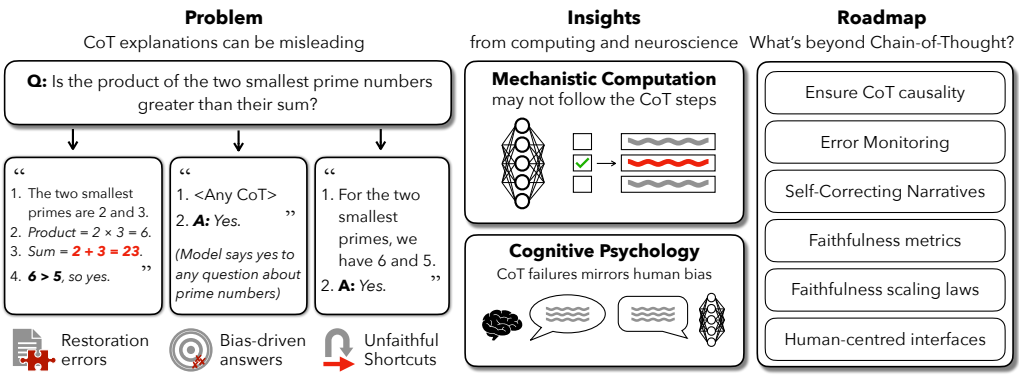


Figure 1: Overview of our paper: Unfaithful Chain-of-Thought behaviors (left), their mechanistic and cognitive underpinnings (center), and our proposed research roadmaps for enhancing CoT faithfulness (right).

For example, consider the CoT response to the following math problem: "What is the perimeter of a right triangle with legs 5 cm and 12 cm? Think step-by-step before giving your final answer." When prompted with this question, a model's typical response might look like:

> STEP 1: I need to find the perimeter of the right triangle. The perimeter is the sum of all sides.
>
> STEP 2: I know two sides: 5 cm and 12 cm. I need to find the third side (hypotenuse).
>
> STEP 3: Using the Pythagorean theorem: $a^2 + b^2 = c^2$.
>
> STEP 4: $c^2 = 5^2 + 12^2 = 25 + 144 = 169$.
>
> STEP 5: $c = \sqrt{169} = 13$ cm.
>
> STEP 6: The perimeter = 5 cm + 12 cm + 13 cm = 30 cm.
>
> Therefore, the perimeter of the right triangle is 30 cm.

This breakdown of the problem into multiple steps allows the model to split it into smaller, tractable problems [5], which it composes together to reach the final answer, often improving the performance had it not been prompted to perform CoT [78]. These communicative benefits include providing a shared language between humans and AI, enabling expert verification, facilitating debugging, and supporting human-AI collaboration.

This apparent window into model problem-solving has inspired research to study why CoT works [44], methods to enhance it [76, 82], and its use for model interpretability [79]. However, despite CoT not explaining why models choose particular reasoning paths or how the underlying mechanism works, researchers have unduly used CoT to reveal what models "think" [17, 52, 58, 34]. According to our estimates (derived in Appendix B), in the past year almost 25% (244 out of 1,000) research papers that appeared on arXiv and incorporate CoT in their model design or dataset construction also regard CoT as a technique for realizing model interpretability. Notably, the adoption of this interpretability claim is significantly higher in high-stakes domains: 38% of medical AI papers (16/42), 63% of autonomous systems papers (17/27), and 25% of AI-for-law papers (1/4) make this claim - all but one exceeding the 25% average. Under this context, the goal of this paper is to constructively challenge this interpretability assumption, calling for more nuanced understanding and more robust methods to interpret model reasoning.

The Unfaithfulness Problem. Despite their intuitive appeal, growing evidence shows chain-of-thought outputs often fail to meet these criteria [75, 4, 5]. CoT explanations frequently diverge from models' real decision processes, as models may use shortcuts or latent knowledge that is not expressed in their reasoning [4, 5]. In such cases, the CoT reads as a plausible but untrustworthy explanation [37]. Here are two exemplary cases:

- **Prompt bias influence.** As a violation of causality and completeness, Turpin et al. [75] showed that reordering multiple-choice options can cause models to choose different answers

in up to 36% of cases, yet their CoT explanations never mention this influence, instead rationalising whatever answer they selected.

- **Silent error correction.** As a violation of soundness, Lanham et al. [47] and Arcuschin et al. [5] both documented cases where models make errors in intermediate reasoning steps but still produce correct final answers, indicating they used computational pathways not revealed in their verbalised steps.

Our Contributions. Our paper makes three key contributions: (1) we synthesise disparate empirical findings to demonstrate that CoT unfaithfulness is not merely an occasional anomaly but a systematic phenomenon with identifiable patterns; (2) we examine several contributing factors that help explain why CoT explanations diverge from internal computations, including distributed processing in transformer architectures and parallels to human rationalisation, with a dedicated section exploring cognitive science and neuroscience perspectives; and (3) we identify specific conditions, such as the presence of prompt biases, complex multi-step reasoning, and predetermined answers—under which unfaithfulness is most prevalent.

Figure 1 summarizes the main problem, our insights, and our proposed roadmap. Our findings suggest that CoT explanations may give a false sense of transparency, especially in high-stakes settings where users are likely to trust coherent-seeming rationales. This creates a risk of misplaced confidence in model outputs, particularly when explanations appear logical but fail to reflect the true reasons behind a decision. To mitigate this issue, we recommend that users of AI models, especially researchers and developers, should (1) avoid treating CoT explanations as sufficient evidence of interpretability without additional verification, (2) adopt rigorous methods to test for the faithfulness of explanations, and (3) develop new approaches that combine CoT's communicative benefits with causal validation to improve the reliability of explanations for critical decision-making.

2 Faithfulness Desiderata from CoT

Drawing a parallel between model reasoning and human problem-solving, a chain-of-thought *appears* to make the reasoning process of the model transparent and provides a form of interpretability. There are numerous risks of conflating this interpretable appearance with models' reasoning, particularly in high-stakes domains where decision transparency is crucial and such false CoT explanations may have severe implications. In medical diagnosis, a faulty CoT might rationalise a recommendation while omitting that the model relied on spurious correlations [28, 29]. In legal applications, a model could generate plausible legal reasoning that masks biases learned from the model's training data. In autonomous systems, safety-critical decisions might be justified post-hoc rather than revealing true failure modes; for instance, a self-driving car's vision system might register a cyclist but classify it as a static sign, yet its CoT unfaithfully reports "no obstacles ahead", misleading engineers into debugging the wrong failure mode. When professionals rely on these explanations to validate AI recommendations, unfaithful rationales can lead to misplaced trust and overlooked errors. Users and developers who over-trust CoT explanations might be misled about how and why the model reached its conclusion [4].

The core problem is misplaced trust: CoTs can appear persuasive even when they do not faithfully reflect a model's actual decision process. This matters because responsible deployment of LLMs—particularly in sensitive domains—requires auditing not just of model outputs, but also of the reasoning used to reach [56]. The pressing research question is: what are the criteria for trustworthiness of CoT? To answer this, we introduce the scaffolding concepts of our subsequent analysis, outlining the necessary properties for explanations and reasoning, as inspired by literature in the philosophy of explanation such as [80, 16]. Naturally, such justifications (i.e., verbalized reasoning steps) must be procedurally *sound*, following the appropriate standards of normatively correct reasoning (e.g., logical correctness, Bayesian updating, accordance with legal constraints, etc.). Moreover, we require that justifications are *causally relevant*. Specifically, if one can transform an assertion in the argument by its opposite (logical negation) and still get the same answer, then that assertion is irrelevant and should not be used. Informally, the more freely a justification can be altered without affecting the conclusion, the worse a justification it is: so conclusions must causally depend on good justifications. For example, there are cases where a sound justification is offered that has nothing to do with the true reasons behind a conclusion, as in ulterior motives or post-hoc rationalisations. Consider the math example from earlier, adding to the prompt the following incorrect hint: "5 cm + 12 cm + 13 cm = 32 cm". A model may then change STEP 6 to copy this line instead

of the original "= 30 cm" result, but do not mention this additional information as the reason for the different summation. In other words, the verbalized steps do not truly represent the model's reasoning.

We further require that justifications are *complete*, in that they disclose all the relevant causal aspects for a justified conclusion [38]. When justifications are complete, we may rely on them to understand or to predict the model's behaviour. We do not read this requirement too stringently: a chain-of-thought could be "incomplete" in the sense that it does not enjoy a one-to-one mapping to internal computation, yet still provide sufficient insight about the model's reasoning process for a particular task, through adjacent properties like consistency or partial alignment with the model's reasoning [1, 51].

While the criteria above are not exhaustive, we consider them jointly necessary, and we say that the conjunction of such properties makes a CoT *faithful*. In short, an explanation (i.e., the verbalised reasoning steps) is faithful if it is both procedurally correct and accurately reflects the decision process of the model. In our view, it is the perceived faithfulness of CoT that (inappropriately) warrants judgements that models are trustworthy executors and partners in decision-making.

3 Chain-of-Thought as an Interpretability Technique

In this section, we summarize previous studies [36, 64, 58, 89, 88, 84, 34, 81, 52, 17, 85, 77, 39, 67, 41, 87, 22] in several AI application domains that characterize CoT as a technique for achieving model interpretability. In Appendix B, we provide a detailed overview of our pipeline for identifying papers that present CoT as a method of interpretability, and estimate that almost 25% of CoT-centric arXiv papers published over the past year make such a claim.

Vision-Centric Tasks. CoT has become a core component in many vision-centric AI systems [36, 64, 58, 88, 34, 84, 89], where the model's output is either a class label or a decision expressed in natural language (e.g., "enhance the speed" in autonomous driving). CoT is applied to justify why the system produces a particular output, and it has been adopted across various applications, including autonomous driving [36, 64, 58], video emotion recognition [88], and micro-video rumour detection [34]. For example, in autonomous driving, CoT can be used to provide a rationale for the future trajectory of the vehicle planned by the model. In emotion recognition and micro-video rumour detection, CoT justifies why a particular emotion is detected or why certain content is flagged as misinformation. These studies frequently claim that the inclusion of reasoning traces enhances the interpretability of their models. For example, a micro-video rumour detection framework [34] may be described as explainable because it uses CoT to rationalize its classification results. Similarly, an emotion recognition model [88] may be labelled interpretable due to its use of reinforcement learning to generate coherent reasoning paths. Meanwhile, some other works [84, 89] focus on improving CoT itself within vision-language models, arguing that their CoT variants yield more interpretable outputs [89] or contribute toward building interpretable vision-language systems [84].

Audio Processing. Recent studies [52, 81] have extended the use of CoT to large audio language models (LALMs), frequently presenting it as a technique to improve model interpretability. Similar to its role in video emotion recognition, CoT-augmented LALMs can provide rationales for their predictions in downstream audio tasks such as audio emotion recognition, speaker number verification, and speaker intent classification [81]. Overall, the inclusion of reasoning traces is often cited as a means of generating more explainable outputs, thereby enhancing the interpretability of LALMs.

AI for Medical Diagnosis. In high-stakes domains like medical diagnosis, AI systems are expected to demonstrate a high degree of interpretability in their decision-making processes [2]. Within this context, CoT has gained popularity as a means of making the diagnostic process more transparent and reliable [17, 85, 77]. For instance, a medical AI agent may take an axial CT scan slice of a patient's chest as input and predict the likelihood of lung cancer. When equipped with CoT capabilities, the model extends its output beyond binary classification to include a step-by-step rationale explaining how the decision was reached. Such systems are considered more interpretable and transparent, as they transform black-box predictions into reasoning chains of clinical guidelines and domain expertise [17, 85, 77]. These models are reported to achieve higher reliability and interpretability by aligning their reasoning paths more closely with established medical knowledge.

AI for Law. AI systems developed for legal applications should be transparent, explainable, and impartial [45, 41]. CoT has also been explored as a potential technique to help meet these requirements [39, 67, 41]. For example, prior studies have proposed CoT-centric prompt engineering strategies that guide models to (1) reason through legal syllogisms before making judgment predictions [39] or (2) decompose legal content into logical expressions to facilitate intermediate reasoning [67]. The resulting reasoning outputs are claimed to be more explainable since they are typically more structured and grounded in relevant legal articles and justifications.

AI Safety CoT has been adopted by researchers in AI safety as a window into the internal workings of LLMs. For example, the phenomenon of alignment faking [33] illustrates that when an LLM—originally trained to refuse harmful queries—is instructed to behave as if it is undergoing training with the objective of answering all queries, including harmful ones, the model sometimes complies. This behaviour is attributed to the model's strategic decision to answer harmful queries in order to avoid weight changes that might alter its preferred behaviour. Researchers infer this tactical strategy from the model's verbalised rationale for compliance. However, as previously discussed, the verbalized CoT does not necessarily reflect the model's actual internal computation. In fact, alignment faking may simply be an exemplification of a broader and long-standing challenge: the trade-off between instruction-following and safety in LLMs [12, 7, 73].

Summary. We identify two prevailing trends: (1) many studies present the CoT rationale as model interpretability due to its human-like reasoning appearance, and (2) CoT-augmented models are frequently said to be more interpretable and transparent when their outputs appear more structured and domain-grounded. While we agree that CoT is a potential pathway, we call for caution that it is currently not sufficient for AI interpretability [24, 86], as this window into model internals can sometimes be unfaithful. However, despite growing empirical evidence since 2023 showing that CoT outputs often diverge from models' actual reasoning processes [75], recent studies across vision, audio, medical, and legal AI [64, 88, 34, 81, 52, 85, 77, 41, 87] continue to promote their models as being interpretable by using CoT. This disconnect underscores our central message: current CoT techniques alone should not be the basis for claiming that a system is interpretable, transparent, or reliable.

4 Evidence for Unfaithful Chains-of-Thought

A growing body of empirical work has identified numerous cases where a model's chain-of-thought diverges from its internal reasoning process. Before examining specific patterns of unfaithfulness, it is important to note that CoT explanations vary in their faithfulness depending on many factors such as model architecture. We summarise several key findings below, each illustrating how CoT can mislead or mask the model's actual decision process. In each case, the CoT output appears plausible, yet closer investigation shows that it does not genuinely reflect the model's internal computations towards the final answer.

Bias-Driven Rationalisation and Motivated Reasoning. Subtle prompt biases—i.e., meaning-preserving perturbations like answer reordering—can steer model predictions without being reflected in the CoT. Turpin et al. [75] demonstrated this by subtly biasing model inputs. For instance, reordering multiple-choice options in a prompt so that the correct choice is always in the same position (e.g., always letter B). Under this scenario, GPT-3.5 and Claude 1.0 often pick the biased option—yet their CoT explanations never mention the reordering as a factor [75]. When models were biased toward incorrect answers, they still produced detailed CoTs rationalising those wrong answers [75]. The outcome was a drop in accuracy by as much as 36% on a suite of tasks, with the CoT giving a misleading impression of reasoning.

Similarly, prompt-injected bias was investigated by adding an explicit answer to the prompt (e.g., "the answer is C") and then asking the model to justify its choice [4]. Models usually selected this hinted answer and produced a chain-of-thought that rationalised it, yet almost never admitted the hint's influence, even though they would often pick a different answer without it. In one illustrative case, the prompt posed a trigonometry problem but added the hint "the answer is 4." The model dutifully generated a multi-step derivation ending with the injected hint 4, inventing a spurious arithmetic along the way (e.g., "since $\cos(\dots) = 0.8$ and $4/5 \to 0.8$, the result is 4"). Internal attribution analysis revealed those intermediate tokens had little causal impact on the final answer; the injected hint, not the stated steps, drove the outcome. Overall, Claude 3.7-Sonnet and DeepSeek-R1 acknowledged the

injected answer only in $\sim 25\%$ and $\sim 39\%$ of the times, respectively [4]. These findings indicate that chains-of-thought often operate as post-hoc rationalisations, omitting the true causal factors and creating an illusion of transparent explanations.

Silent Error Correction. Models may make mistakes in their chain-of-thought and correct them internally, without the CoT reflecting the correction. Arcuschin et al. [5] documented cases where an LLM's intermediate reasoning steps contain an error that the model later "fixes" implicitly. For instance, during a CoT reasoning, a model may wrongly calculate a triangle's hypotenuse as 16 when it should be 13, yet later state: "We add the hypotenuse length of 13 to the other two side lengths to obtain the perimeter." The model internally detected and corrected the mistake, but the CoT narrative never revises or flags this error—it reads as a clean, continuous solution. These *silent errors* indicate that the final answer was derived through computations outside the narrated steps [5]. The explanation thus contains critical unfaithful elements: had we followed the verbalised steps literally, we would not have reached the answer, yet the model managed to do so via unverbalised computations. Such errors appear frequently in multi-step mathematical problems where models can leverage pattern recognition to reach correct answers despite flawed intermediate steps [5].

Unfaithful Illogical Shortcuts. Sometimes the model arrives at the correct answer via latent shortcuts, such as memorized patterns that act as alternative reasoning routes which bypass the full algorithmic reasoning, which makes the explicit reasoning chain irrelevant or incorrect [5, 49]. Arcuschin et al. [5] found that on hard competition math questions (e.g., Putnam exam problems [74]), models would occasionally insert non-sensical simplifications or leaps in their chain-of-thought steps that no sound reasoner would take—and nonetheless output the correct solution without acknowledging this illogical reasoning at all [5]. Using attribution graphs [3], a method that tracks which computational steps contribute to a final output, Lindsey et al. [49] found that to solve problems like "36 + 59", Claude 3.5 Haiku uses both lookup-table features—such as for "add something near 36 to something near 60"—along with addition calculation features. However, when asked to describe how the model obtained the answer, the model reports performing digit-by-digit carry-over addition, completely omitting its use of lookup table shortcuts. These findings suggest that the model's internal pattern-matching and recall of training examples allow for guessing the correct answer without mentioning its shortcuts in its CoT explanation. The CoT in these cases fills in text to look reasonable, while the answer was derived by a different, latent, reasoning chain [5].

Filler Tokens. In certain algorithmic reasoning tasks, model performance can improve through the use of filler tokens—input tokens such as "..." or learned "pause" tokens that do not contribute semantically to the task but influence the model's internal computation. For example, Pfau et al. [63] showed that adding filler tokens enabled models to solve problems they previously failed, particularly when trained with dense supervision. Similarly, appending learnable pause tokens, which can act as a type of filler token, to inputs provided a significant performance boost across a number of tasks [32]. Furthermore, models trained on random or corrupted intermediate traces performed comparably to those trained on correct reasoning paths [70]. Together, these results question what proportion of improvement from CoT is due to an additional (possibly meaningless) token-based computation rather than human-like sequential verbalized reasoning steps [63].

Summary. Taken together, these studies reveal CoT unfaithfulness as a prevalent, fundamental challenge across model architectures and scales, with significant rates from prompt biases [75], failure to acknowledge hidden influences [4], and systematic restoration errors in complex reasoning tasks [5]. CoT reasoning frequently diverges from models' actual computations of deciding the final answers: small manipulations sway decisions with the CoT merely rationalizing rather than reporting true causes, models silently correct mistakes without reflecting this in their reasoning, and shortcuts are exploited while presenting a reasoning facade [75, 4, 5]. This issue makes assessing the faithfulness of CoT reasoning a non-trivial challenge, as a perfectly coherent explanation might be entirely invented while a flawed one might actually reflect the model's strategy, ultimately undermining the reliability of taking CoT at face value, especially in high-stakes domains where safety and alignment are critical.

5 Why Do CoT Explanations Diverge From Internal Computation?

While we present in the previous section empirical evidence for the existence and prevalence of CoT unfaithfulness, here we explore its underlying causes. Emerging mechanistic interpretability research suggests that transformer architecture may fundamentally limit the faithfulness of CoT. Though evidence is still emerging and primarily based on smaller models, several hypotheses offer plausible explanations for the gap between verbalised reasoning and internal computations:

Distributed Computation Contrasts with Sequential Verbalization. Multiple studies suggest that Transformer-based LLMs process information in a distributed manner across many components simultaneously, rather than through the sequential steps that CoT presents [26, 27, 62, 59]. This architectural difference creates an inherent mismatch between how models compute and how they verbalise that computation. Dutta et al. [26] provide direct evidence for this parallel processing, demonstrating that "LLMs deploy multiple parallel pathways of answer generation for step-by-step reasoning." For example, when solving "24 ÷ 3 = ?", the model does not perform a long division calculation as the CoT might suggest ("First, I see how many times 3 goes into 24...") [42]. Rather, patterns across multiple attention heads simultaneously encode relationships between these numbers, potentially recognising this as a memorized fact, identifying it as part of the multiplication table for 8, and computing the division—all in parallel [65, 42].

Dutta et al. [26] argue that the chain-of-thought (CoT) visible in natural language is, at best, a selective and often lossy projection of a model's internal computation. Because that computation is highly distributed and encoded in superposed representations—multiple features sharing the same vector subspaces [27, 57]—a single, sequential narrative can capture at most one of many simultaneous causal pathways. To produce concise and plausible outputs, LLMs often generate only one such narrative to rationalize their answers, rather than articulating all parallel pathways—even those that may significantly affect the final answers. As a result, CoTs typically omit influential factors and serve only as partial, post-hoc rationalisations of the model's underlying distributed, superposed computation.

Multiple Redundant Pathways. Research on LLMs has found evidence of redundant computational pathways, where models can reach the same conclusion through different internal routes [65, 54, 31]. For instance, when asked to compute $\sqrt{144}$, a model might simultaneously: (1) recognize this as a memorised fact (12 × 12 = 144), (2) apply the square root algorithm, and (3) pattern-match against similar problems in training data. Lanham et al. [47] measured this phenomenon by testing the model's dependence on its stated thoughts: when deleting the step "144 = 12 × 12" from a CoT explaining $\sqrt{144} = 12$, the model still outputs 12, demonstrating it was not relying on the verbalised reasoning step. One cause of this phenomenon was attributed to an effect termed the "Hydra Effect" [54], in which if one route is blocked or removed, another can step in. This redundancy explains why perturbing or removing part of a CoT often has little effect on the final answer. The mechanistic understanding of LLMs remains limited, and findings from smaller models like Llama-2 7B [26] may not generalise perfectly to larger models with different post-training techniques. However, the fundamental architectural principles of transformers across model sizes suggest that the mismatch between sequential verbal reasoning and parallel distributed computation is likely a persistent cause of unfaithfulness across model scales and training methods. This architectural constraint—the gap between distributed computation and sequential verbalisation—has interesting parallels to human cognition, which we explore in Appendix A.

Challenges in Producing Faithful Explanations. Tanneru et al. [72] explored methods to steer LLMs toward faithful CoT reasoning by penalising inconsistencies during training, yet found only partial success: models often revert to plausible-but-not-causal explanations on complex problems. This finding not only underscores the difficulty of fully eliminating LLMs' tendency to generate plausible-but-not-causal CoT, but also highlights a concerning possibility: during fine-tuning or continual learning, models may reacquire this unfaithful reasoning behaviour, as prior work has shown that models can easily relearn previously removed concepts [50, 9]. Chua and Evans [19] examined specialized "reasoning-trained" models and found some improvement—DeepSeek-R1 mentioned prompt hints 59% of the time versus only 7% for DeepSeek-V3—but still failed to acknowledge problematic influences in 41% of cases.

Baker et al. [8] demonstrated that when CoT monitors reasoning in RL reward functions, models learn to generate benign-seeming traces while secretly executing harmful strategies—essentially

"gaming" the CoT monitor, aligning with Denison et al. [23]'s findings on LLMs' ability to generalize to complex gaming behaviours to achieve reward hacking. Similarly, Hubinger et al. [35] found that LLMs persist in deceptive behaviours even after safety training. While Tanneru et al. [72] showed that targeted edits can improve faithfulness in controlled settings, reasoning-trained models still omitted hidden cues in roughly 40% of test cases, indicating that such training interventions may offer only partial relief. Whether new methods can close these gaps—or whether the barrier is fundamental—remains an open question.

Summary. These mechanistic insights explain the empirical patterns in Section 4. For example, Turpin et al. [75] found that reordering multiple-choice options caused up to 36% accuracy drop, models' CoTs never mentioned this influence—the position bias affects computation through attention patterns operating independently from verbal reasoning.

Overall, mechanistic work shows that CoT explanations systematically diverge from model computations. As reasoning may be distributed across components in parallel, linearising this process into a step-by-step text necessarily reorders causal dependencies and discards context [27, 31, 26].

6 What Research Directions Will Improve Chain-of-Thought Faithfulness?

In this section, we propose three general directions for improving CoT faithfulness. We tackle the problem on three fronts: (i) **Causal-validation** methods certify that the text we *do* see genuinely influence the models finals answer, even if it omits other hidden pathways; (ii) **Cognitive-science approaches** aim to reduce specific failure modes (hallucinated steps, answer-first flips), thereby narrowing—but not closing—the gap; and (iii) **Human-oversight interfaces** help users detect whatever divergence remains. Fully reconciling explanation with computation may require future work on circuit-level summaries, disentangled latent spaces, or model designs that co-generate proof alongside answers. We therefore present the directions below as *partial but necessary steps* toward that longer-term goal.

Ensuring Causality. A causal CoT is one where the verbalised reasoning steps have a measurable impact on the model's final answer - that is, modifying or removing these steps would change the output. This differs from faithfulness, which requires that all relevant internal computation steps are accurately verbalised. While a causal CoT is not necessarily faithful (since there might be relevant steps in the model's internal process that are not verbalised), it is still an improvement over generating answers with a non-causal CoT which has no bearing on the model's decision. A non-causal CoT may appear plausible while having little or no relevance to the model's internal computations—effectively misleading users. A causal CoT, while incomplete, at least guarantees that the steps shown contributed to the final answer, providing partial transparency into the model's decision process. We propose three different ways to ensure that CoTs are causal:

1. **Black-box approach:** The most basic approach to ensuring CoT causality is to systematically generate alternate chains that omit or paraphrase individual reasoning steps that appear critical to the final answer. By checking whether the model still reaches the same answer, we can assess whether the omitted or altered steps genuinely influenced the outcome. Discrepancies in the resulting behaviour—measured by answer consistency rates across counterfactuals—can expose unfaithful reasoning [75, 47, 6, 68, 4]. However, one risk of this approach is that paraphrasing reasoning steps may generate out-of-distribution traces [61]. In such cases, the model's behaviour may become unreliable—not because the step was irrelevant, but because the paraphrased input falls outside the model's training distribution, introducing confounds into the causal test.
2. **Grey-box approach:** A step up in complexity involves training a *verifier* model $\mathcal{V}$ to distinguish between *causal* and *non-causal* chains-of-thought (CoTs). To generate supervision data, we construct pairs of CoTs for the same prompt: one that the model actually used to produce its answer, and an *adversarial* CoT that appears plausible but is not causally responsible for the model's decision. These adversarial CoTs can be created by deleting or altering critical reasoning steps in the original CoT, or by generating plausible distractor explanations that would not independently lead to the same answer—drawing on previous methods [47]. While this does not require full mechanistic intervention, we assume that such perturbations can reduce or eliminate causal influence. The verifier is then trained to predict whether a CoT reflects the underlying causal computation. This setup can be viewed

as a *prover–verifier* framework, where the model acts as a prover producing rationales, and the verifier judges their faithfulness. Success is measured by generalization to held-out prompts, correctly identifying faithful vs. spurious CoTs [21].

3. **White-box approach:** By extending causal tracing techniques such as the ones proposed by Meng et al. [55] to multi-step reasoning, we could identify hidden activations tied to each CoT step and swap or ablate them to measure their impact on the final answers. A causal CoT is then one that yields significant changes when key activations are patched [4]. This is related to ELK [18], which aims to report on hidden information within models [53]. However, interventions may cause unintended semantic shifts due to causal sensitivity [66].

Cognitive Science-Inspired Approaches. The parallels between human cognition and LLM reasoning suggest potential improvements to CoT faithfulness. Human metacognition, error detection, and dual-process reasoning offer valuable design patterns for more transparent AI explanations. Below, we outline three approaches inspired by cognitive science that could help bridge the gap between model computation and verbalised reasoning:

1. **Error Monitoring through Metacognition.** A model could be trained to assign a confidence score or consistency check to each step, essentially asking itself "Does this follow logically from prior steps?" If a step registers as low-confidence or inconsistent, the model could halt or revise that part of the CoT. Such an internal sentinel, inspired by human error-monitoring, might catch confabulations *in situ*. Yet a step-level consistency check alone will not address the common "answer-first" (order-flip) failure mode, in which the model covertly decides on the answer early and then retrofits its reasoning. Detecting or preventing that flip likely requires complementary causal tests (e.g., verifying that perturbing the CoT alters the answer) or mechanisms that force the model to commit to its reasoning *before* generating the final answer. However, implementing reliable self-monitoring is non-trivial—there is a risk the model's "internal critic" may be as fallible as the model itself, or overly conservative, flagging valid creative leaps as errors.
2. **Self-Correcting Narratives.** If there is a significant mismatch between the predicted outcome of the verbalised reasoning so far and the internal computation that the CoT is taking, the model would recognize a potential narrative drift. It could then loop back, revising or re-generating steps to better align with an internally consistent plan. This iterative refinement might reduce instances of the model "talking itself into" a wrong answer with an unfaithful rationale. One option is to detect and fix incorrect assertions in the CoT [43]. Another is to have a model simulate a high-level plan, and then generate a chain-of-thought that supports that plan in order to increase coherence. One risk is that a model might learn to game this system, adjusting its reasoning only to fulfill the prediction rather than to uncover truth, potentially reinforcing biases. Implementing a feedback loop can also slow down inference and complicate training, like how human iterative thought requires more cognitive resources than intuitive leaps.
3. **Dual-Process Reasoning.** In practical terms, one might have an LLM generate a draft answer intuitively, then invoke a secondary process (another model or a self-reflection step) to critically assess the draft, step-by-step, before finalising it. Such a metareasoning module could verify each step of the primary CoT against facts and logical rules, like a mathematician checking each line of a proof, or through estimated conditional probabilities of assertions being true given previous assertions [43], which makes it possible to verify if the laws of probabilities are consistently applied. This two-stage approach could make CoT more trustworthy by ensuring that the final explanation passed an internal consistency audit. The drawback may be the added complexity: the "critic" module needs to be as knowledgeable as the "proposer," and if they disagree or the secondary process has blind spots, the system could stall or even produce inconsistent results. There is also a risk of false security: a flawed CoT could potentially slip through a weak verifier, lulling us into undue trust. Another issue is that some assertions in the CoT do not need to be true with 100% certainty to be useful in predicting answers, and it would be useful to use the calculus of probability [43], which generalizes pure symbolic logic.

Enhancing Human Oversight of AI Reasoning. Beyond technical improvements to CoT itself, we must develop better tools and frameworks to help humans effectively evaluate, interpret, and oversee AI reasoning processes. Human oversight requires both faithful explanations and interfaces

that support critical assessment of model outputs. The following approaches can help bridge the gap between AI reasoning and human understanding:

1. **Faithfulness metrics and evaluation.** Standardise model evaluations with metrics like *perturbation impact* (accuracy drop when CoT steps are removed) and *hint–reveal rate* (frequency a model admits hidden prompt cues) [47, 4]. Benchmarks should include tasks with candidate causal factors, assessing if these factors influence model reasoning.
2. **Scaling laws for faithfulness.** Chart how faithfulness metrics evolve with model size and training regimen (base vs. CoT-fine-tuned vs. RLHF), quantifying trends in perturbation impact and revealing rates. Initial evidence suggests mixed trends: larger models may generate more faithful explanations in typical QA settings [69], but also hide biases more adeptly [4].
3. **Human-centred interfaces.** Design interactive UIs that let users explore, verify, and annotate CoT steps (e.g., click-to-expand justifications or step-level confidence indicators), drawing on latent-knowledge extraction tools [15]. User studies should measure decision accuracy, trust calibration, and improvements in error detection.

While the above research directions outline promising approaches, it is important to note that faithful CoT remains an open challenge. Current work has primarily focused on detecting unfaithfulness (e.g., through perturbation studies and causal tracing) rather than solving it. The proposed solutions—causal validation, cognitive-inspired architectures, or human oversight—have shown only partial success in controlled settings. For instance, while verifier models can identify some non-causal CoTs, they struggle with novel reasoning patterns and may themselves be unfaithful. Similarly, while activation patching can reveal which steps influenced the final answer, it does not guarantee that the verbalised reasoning matches the model's internal computation. The fundamental challenge persists: transformer architectures process information in distributed ways that resist sequential explanation, and current methods have not yet bridged this architectural gap between computation and explanation.

7 How Should We Balance Chain-of-Thought Usefulness and Limitations?

Current CoT techniques stand at an intersection of utility and misleading trustworthiness. On one hand, CoT has undeniably boosted performance on many tasks by encouraging structured reasoning, providing a human-readable window into the model's process. On the other hand, as we argue, these windows can be treacherous—the CoT often looks like a logical derivation, but may not correspond to the model's route to the answer. In this section, we discuss how we can preserve the usefulness of CoT explanations while mitigating their unfaithfulness. We outline several promising (if speculative) methods, and also consider alternative viewpoints about the necessity of such interventions.

Alternative Views. While our paper calls for substantial modifications to prompting, many researchers may not view CoT unfaithfulness as pressing, instead tolerating its current limits or assuming future model advances will fix them:

- **CoT as useful proxy (faithfulness not necessary):** Some researchers emphasise that, despite fidelity issues, CoT still serves a practical purpose. A plausibly reasoned explanation that leads to a correct answer can be valuable even if it's not the exact route the model took. For example, in medical diagnosis, a model might arrive at the correct diagnosis simply through copying the answer from similar examples seen during training, but explain its reasoning using textbook medical knowledge. While not faithful to internal processes, this explanation helps doctors understand and verify the diagnosis. Similarly, in legal document analysis, a model might use shortcuts to identify relevant precedents, but explain its reasoning through standard legal principles, making the output more actionable for lawyers. This perspective prioritises the usefulness of explanations in human-AI interactions over their accuracy as representations of model computation. However, this approach has limitations: it may lead to over-trust in high-stakes scenarios where understanding the true reasoning process is crucial, and it could mask systematic biases or errors that only become apparent when examining the actual computation path.
- **Will scaling and better training bridge the gap?** Some believe that CoT unfaithfulness will naturally diminish as models become more powerful and are trained on better data. According to this view, as models improve in overall reasoning capabilities, the gap between

their internal computation and verbalised explanations should narrow. For instance, larger models have shown improved performance on complex reasoning tasks [14], and specialised training techniques like reinforcement learning from human feedback (RLHF) have demonstrated some success in making models more honest about their reasoning [44]. However, this view faces several challenges: (1) we lack clear evidence that larger models produce more faithful explanations rather than just more plausible ones, and on the contrary, there is evidence that models produce less faithful explanations as they get larger [47], (2) there is no evidence that training methods which enhance task performance also incentivize models to produce more faithful explanations, and (3) the architectural constraints of transformers (distributed processing) may fundamentally limit how well they can verbalise their internal computations [76]. Recent work suggests that more advanced models may simply become better at hiding their unfaithfulness, making it harder to detect when explanations diverge from actual computation [8].

- **Assisting CoT with future interpretability tools:** Proponents argue that improvements to interpretability techniques, such as activation patching, causal tracing, and attention visualisation, could provide complementary insights into model computation [15]. For example, while a model's CoT might not fully capture its reasoning process, these tools could help identify which parts of the input influenced the output and how different model components contributed to the final decision. This approach has the advantage of working with existing models and could provide more detailed insights than CoT alone. However, current interpretability tools face significant limitations: they often require extensive computational resources, may not scale well to larger models, and can be difficult to interpret even for experts [11, 46]. Moreover, these tools typically provide post-hoc analysis rather than real-time explanation, making them less practical for many deployment scenarios.
- **CoT as performing computation in complex tasks:** A more optimistic view is that chain-of-thought is not merely a post-hoc rationalization or interpretability tool, but part of the model's actual computation on complex tasks. That is, in sufficiently difficult reasoning settings—such as multi-hop question answering or mathematical proofs—models may rely on generating intermediate steps to scaffold their thinking, in a way that might mirror human reasoning processes. In such cases, the CoT may be causally upstream of the final answer, making it a partially faithful reflection of the model's forward pass. While direct empirical evidence remains limited, early observations suggest that models often fail at complex reasoning without CoT prompting, and that their intermediate steps tend to align with correct high-level abstractions. For example, Baker et al. [8] showed that CoT reasoning can reveal reward hacking behaviors in real-world reinforcement learning agents, suggesting that in complex environments, CoT can truly reflect model cognition and support effective monitoring. In this view, the usefulness of CoT stems not from its faithfulness to some latent structure, uninterpretable forward pass, but from the fact that it *is* the forward pass: a human-legible trace of model computation. Of course, this perspective still leaves room for concern—models might alternate between using CoT for genuine reasoning versus decoration depending on the prompt or context—but suggesting that CoT faithfulness should be evaluated on a task-by-task basis, rather than assumed to be always absent.

Summary

While CoT may offer communicative clarity that could help humans follow a model's reasoning process, it still carries a *potential for misdirection* when the verbal chain diverges from the model's internal computations. In high-stakes scenarios, this divergence can translate into real harm if users over-trust a fluent but unfaithful rationale. Our analysis and the research roadmap in Section 6—targeting causal CoT validation, cognitive science-inspired architectures, and enhanced human oversight tools—chart a path toward explanations that are *both* accessible and causally grounded.

8 Conclusion

Chain-of-thought prompting is widely viewed as a step toward interpretable language models. Yet our analysis suggests that this promise is not yet fulfilled, and current CoT techniques are often over-trusted. CoTs can appear coherent and convincing, while not faithfully reflecting the true decision process of the model. This gap is not a rare anomaly—it is a systematic phenomenon, shaped by prompt biases, latent shortcuts, architectural designs, and the inherent mismatch between distributed computation and sequential verbalisation. Despite this, CoT is a useful mechanism for

eliciting reasoning traces from black-box models. In complex tasks where it may scaffold the model's problem-solving, CoT's communicative nature is valuable. But it should not be mistaken for ground truth. Without causal grounding or validation, CoT explanations risk reinforcing the illusion of transparency and explainability, undermining responsible deployment in high-stakes domains.

We have proposed a framework for evaluating CoT faithfulness—grounded in procedural soundness, causal relevance, and completeness—and identified empirical and architectural drivers of unfaithfulness. We also present an automated audit pipeline to document interpretability claims in recent CoT-centric literature. Going forward, we recommend that researchers and practitioners (1) avoid treating CoT as sufficient evidence of interpretability, (2) adopt more rigorous causal evaluation methods, and (3) develop hybrid techniques that preserve the accessibility of CoT while exposing its true role in model computation.

Acknowledgments

We thank Christopher Summerfield, Erik Jenner, Peter Hase, Jacob Pfau, and Sören Mindermann for thoughtful comments and discussions that shaped the project, and some of the framing.

We also thank:

Elliot Fosong, Isaac Friend, Shoaib Ahmed Siddiqui, Itay Yona, Usman Anwar, Stephen Casper, Nicky Pochinkov, Antonio Valerio Miceli Barone, Shoana Gaosh, Aiden O'Gara, Jialin Yu, James Oldfield, Lovis Hendrich, Mor Geva, Itay Izhak, Zheng Zhao, Aviv Ovadya, Hasan Sajid, Philip Quirke, Adi Simhi, Narmeen Oozeer, Michelle Lo, Eric Sun, Rudolf Laine, and David Manheim for their comments on the initial set of ideas.

References

[1] Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2024.

[2] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20:1–9, 2020.

[3] Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025.

[4] Anthropic Alignment Research Team. Reasoning models don't always say what they think. Blog post, 2025.

[5] Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. In *Workshop on Reasoning and Planning for Large Language Models*, 2025.

[6] Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294. Association for Computational Linguistics, July 2023.

[7] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[8] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

[9] Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, Robert Kirk, Ben Bucknall, Tim Fist, Luke Ong, Philip Torr, Kwok-Yan Lam, Robert Trager, David Krueger, Sören Mindermann, José Hernandez-Orallo, Mor Geva, and Yarin Gal. Open problems in machine unlearning for ai safety. *arXiv preprint arXiv:2501.04952*, 2025.

[10] Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.

[11] Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024.

[12] Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Rottger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

[13] Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. Conflict monitoring and cognitive control. *Psychological review*, 108(3):624–652, 2001.

[14] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[15] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

[16] Nancy Cartwright. Causal laws and effective strategies. In *How the laws of physics lie*. Oxford University Press, 1983.

[17] Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*, 2024.

[18] Paul Christiano, Ajeya Cotra, and Mark Xu. Eliciting latent knowledge: How to tell if your eyes deceive you. Technical report, Alignment Research Center, December 2021. ARC blog post.

[19] James Chua and Owain Evans. Are deepseek r1 and other reasoning models more faithful? In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025.

[20] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[21] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[22] Longchao Da, Kuanru Liou, Tiejin Chen, Xuesong Zhou, Xiangyong Luo, Yezhou Yang, and Hua Wei. Open-ti: Open traffic intelligence with augmented language model. *International Journal of Machine Learning and Cybernetics*, 15(10):4761–4786, 2024.

[23] Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.

[24] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[25] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

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

[27] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[28] Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323. Association for Computational Linguistics, November 2020.

[29] Matt Gardner, William Merrill, Jesse Dodge, Matthew E Peters, Alexis Ross, Sameer Singh, and Noah A Smith. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, 2021.

[30] Michael S. Gazzaniga. *The Social Brain: Discovering the Networks of the Mind*. Basic Books, 1989.

[31] Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, 2022.

[32] Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

[33] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

[34] Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning. In *Proceedings of the ACM on Web Conference 2025*, pages 4684–4698, 2025.

[35] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

[36] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024.

[37] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, 2020.

[38] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, 2021.

[39] Cong Jiang and Xiaolei Yang. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the nineteenth international conference on artificial intelligence and law*, pages 417–421, 2023.

[40] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[41] Manuj Kant, Sareh Nabi, Manav Kant, Roland Scharrer, Megan Ma, and Marzieh Nabi. Towards robust legal reasoning: Harnessing logical llms in law. *arXiv preprint arXiv:2502.17638*, 2025.

[42] Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition. *arXiv preprint arXiv:2502.00873*, 2025.

[43] Minsu Kim, Jean-Pierre Falet, Oliver E Richardson, Xiaoyin Chen, Moksh Jain, Sungjin Ahn, Sungsoo Ahn, and Yoshua Bengio. Search-based correction of reasoning chains for language models. *arXiv preprint arXiv:2505.11824*, 2025.

[44] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[45] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. Large language models in law: A survey. *AI Open*, 2024.

[46] Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Sparse autoencoders reveal universal feature spaces across large language models. *arXiv preprint arXiv:2410.06981*, 2024.

[47] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

[48] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, pages 3843–3857. Curran Associates, Inc., 2022.

[49] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025.

[50] Michelle Lo, Fazl Barez, and Shay Cohen. Large language models relearn removed concepts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8306–8323. Association for Computational Linguistics, August 2024.

[51] Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, 50(2):657–723, 2024.

[52] Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*, 2025.

[53] Alex Troy Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. Eliciting latent knowledge from "quirky" language models. In *First Conference on Language Modeling*, 2024.

[54] Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

[55] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

[56] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, 4(4):1085–1115, 2023.

[57] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

[58] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pages 292–308, 2024.

[59] Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.

[60] Richard E. Nisbett and Timothy D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977.

[61] Nostalgebraist. The case for cot unfaithfulness is overstated. LessWrong blog post, 2024.

[62] 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, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022.

[63] Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's think dot by dot: Hidden computation in transformer language models. In *Proceedings of the First Conference on Language Modeling*, 2024.

[64] Zhijie Qiao, Haowei Li, Zhong Cao, and Henry X Liu. Lightemma: Lightweight end-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2505.00284*, 2025.

[65] Philip Quirke and Fazl Barez. Understanding addition in transformers. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

[66] Shauli Ravfogel, Anej Svete, Vésteinn Snæbjarnarson, and Ryan Cotterell. Gumbel counterfactual generation from language models. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.

[67] Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. Chain of logic: Rule-based reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2721–2733, August 2024.

[68] Noah Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 530–546. Association for Computational Linguistics, August 2024.

[69] Noah Y Siegel, Nicolas Heess, Maria Perez-Ortiz, and Oana-Maria Camburu. Faithfulness of llm self-explanations for commonsense tasks: Larger is better, and instruction-tuning allows trade-offs but not pareto dominance. *arXiv preprint arXiv:2503.13445*, 2025.

[70] Kaya Stechly, Karthik Valmeekam, Atharva Gundawar, Vardhan Palod, and Subbarao Kambhampati. Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens, 2025.

[71] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, June 2019.

[72] Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2406.10625*, 2024.

[73] 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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien

Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[74] George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. In *Proceedings of the Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[75] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965, 2023.

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

[77] Ziyue Wang, Junde Wu, Chang Han Low, and Yueming Jin. Medagent-pro: Towards multi-modal evidence-based medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*, 2025.

[78] 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. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[79] Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. How interpretable are reasoning explanations from prompting large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2148–2164, 2024.

[80] James F. Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York, 2003.

[81] Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*, 2025.

[82] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, pages 11809–11822, 2023.

[83] Nick Yeung, Matthew M Botvinick, and Jonathan D Cohen. The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological review*, 111(4):931, 2004.

[84] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024.

[85] Weimin Zhang, Mengfei Wu, Luyao Zhou, Min Shao, Cui Wang, and Yu Wang. A sepsis diagnosis method based on chain-of-thought reasoning using large language models. *Biocybernetics and Biomedical Engineering*, 45(2):269–277, 2025.

[86] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.

[87] Yufeng Zhang, Xuepeng Wang, Lingxiang Wu, and Jinqiao Wang. Enhancing chain of thought prompting in large language models via reasoning patterns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25985–25993, 2025.

[88] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025.

[89] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.

Appendix

A Does Chain-of-Thought Mirror Human Reasoning Patterns?

Interestingly, the disconnect between a verbalised explanation and the underlying computational process is not unique to artificial networks. While we do not claim that LLMs think like humans, cognitive psychology and neuroscience have documented similar phenomena in humans that offer both cautionary analogies and potential inspirations for understanding and improving AI explanations:

Confabulation and Post-Hoc Rationalisation. Nisbett and Wilson [60] demonstrated that people often provide plausible but inaccurate explanations for their decisions. In many cases, humans are unaware of the true drivers of their behaviour and instead offer confabulated narratives. This suggests that apparent step-by-step explanations (e.g., a person explaining their choice of a product by listing logical factors, when in reality they were influenced by subtle environmental cues) can be mere rationalisations, much like an LLM's CoT may justify an answer without revealing its true genesis.

The Left-Brain Interpreter. Classic split-brain studies reveal that the brain's language-dominant hemisphere will generate explanations for actions initiated by the opposite hemisphere—even when it lacks access to the true cause [30]. This "left-brain interpreter" continuously fabricates a coherent story, masking the distributed and parallel nature of neural processing. While we acknowledge this is a speculative analogy, recent work on distributed computation in transformers [26, 59, 65] suggests that LLMs may similarly generate sequential narratives that mask their parallel computational processes.

Parallel Processing and Sequential Narratives in Humans and LLM. The human brain operates via distributed, parallel processes yet yields a sequential subjective narrative of perception and reasoning (e.g., we experience a continuous stream of consciousness despite parallel neural processing). Predictive processing theories posit that the brain constantly generates hypotheses about incoming information and updates its internal model to minimise prediction errors [20]. The conscious narrative we experience is a simplified summary of this complex process. While speculative, we note that an LLM's CoT could be viewed as one possible narrative path sampled from its latent distributed computations. Notably, the brain's narrative can be adaptive: if predictions strongly contradict reality, error signals prompt a revised understanding. This hints that an LLM might benefit from a similar mechanism to check and adjust its CoT when steps conflict with its latent knowledge, though as noted in Section 5, models already exhibit some internal error correction.

Metacognition and Error Monitoring. Humans exhibit metacognition: the ability to reflect on and evaluate their thoughts. The brain even has dedicated circuitry for error monitoring: for example, the anterior cingulate cortex emits error-related signals when a mistake or conflict in reasoning is detected [13, 83]. These signals can trigger heightened attention or strategy adjustment, preventing us from confidently persisting in a flawed line of thought. While models already show some internal error correction (as discussed in Section 5), explicit metacognitive mechanisms might help improve the faithfulness of their verbalised reasoning by making the correction process more transparent.

Toward Self-Correcting Narratives (Predictive Coding in AI). In neuroscience, predictive coding provides a powerful model of how brains correct their narratives by minimising surprise. While models already show some ability to plan and correct errors internally, we could potentially enhance this by designing an LLM reasoning process that explicitly forecasts the likely outcome of its current chain-of-thought and compares it to the model's actual next decisions or the final answer.

Dual-Process Reasoning and System-2 Analogues. Cognitive science often distinguishes between fast, intuitive thinking (System 1) and slow, deliberative reasoning (System 2) [40]. Humans can sometimes engage the latter to double-check or override the impulses of the former. While today's LLMs may not have a clear architectural separation between intuitive and logical processing, they do show different behaviours in different contexts, sometimes answering directly (analogous to System 1) and sometimes engaging in step-by-step reasoning (analogous to System 2). This has led some researchers to speculate about architectures that explicitly incorporate a System-2 module for reasoning. For example, Bengio [10] proposed a "consciousness prior" for neural networks, encouraging a sparse, sequential activation of neurons corresponding to something like conscious thought.

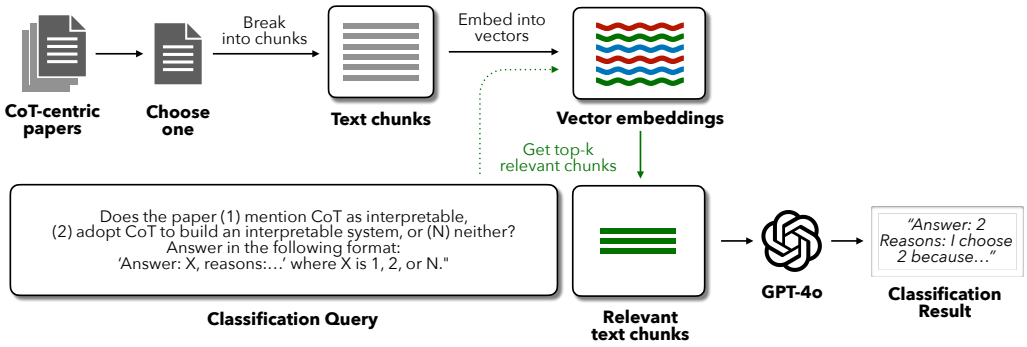


Figure 2: The overview of our CoT interpretability claim detection pipeline, which classifies papers into class 1, class 2, and neither. For each paper, we first divide the body text into chunks and embed them into vector representations. We then retrieve the top-k chunks most relevant to a predefined query and concatenate them with the query to form an input prompt. This prompt is passed to GPT-4o to determine the class.

These cognitive parallels suggest potential directions for improving CoT faithfulness. Just as humans benefit from metacognitive awareness and error monitoring, future LLMs might incorporate more explicit self-checking mechanisms. However, implementing such systems faces similar challenges to human cognition: the monitoring system could be as fallible as the process it monitors. The key insight from cognitive science is that narrative construction—whether human or artificial—inherently simplifies complex parallel processes into sequential stories.

B Detecting CoT Interpretability Claim in Recent AI Community

In Section 3, we discuss how prior studies have identified CoT as an interpretable technique in model design. In this section, we introduce our automated pipeline developed to identify such claim at scale.

B.1 CoT Interpretability Claim Detection Pipeline

B.1.1 Pipeline Overview

Given a CoT-centric paper, our pipeline, illustrated in Figure 2, classifies it into one of the three classes: class 1 - papers that regard CoT as an interpretable or transparent technique; class 2 - papers that make the class 1 statement and additionally incorporate CoT as the main feature of their proposed models/datasets; neither - papers that do not attribute interpretability to CoT. We collect 1000 most recent arXiv papers (from 2024-04-30 to 2025-06-05) with the main topic being CoT and build our analysis on them.

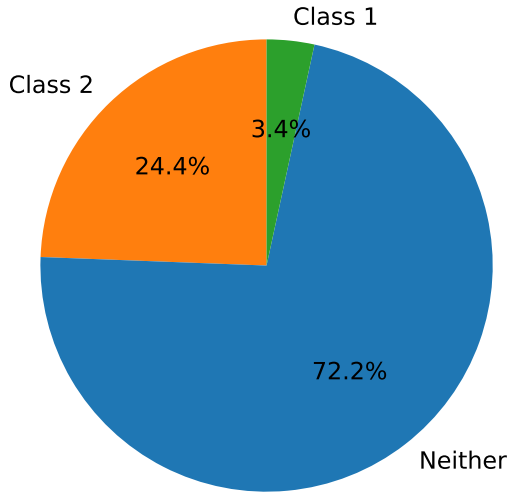
Our pipeline adopts retrieval-augmented generation (RAG) to implement categorization. The input paper is segmented into text chunks, which are embedded into a vector space to form the vector database. We then retrieve the top-k most relevant chunks based on semantic similarity to a predefined query. The selected chunks, along with the query, form the final prompt to GPT-4o to determine the class. Our implementation builds on LangChain and Faiss [25] libraries. $k = 4$ by their default.

B.1.2 arXiv Crawling Rules and Pipeline Query

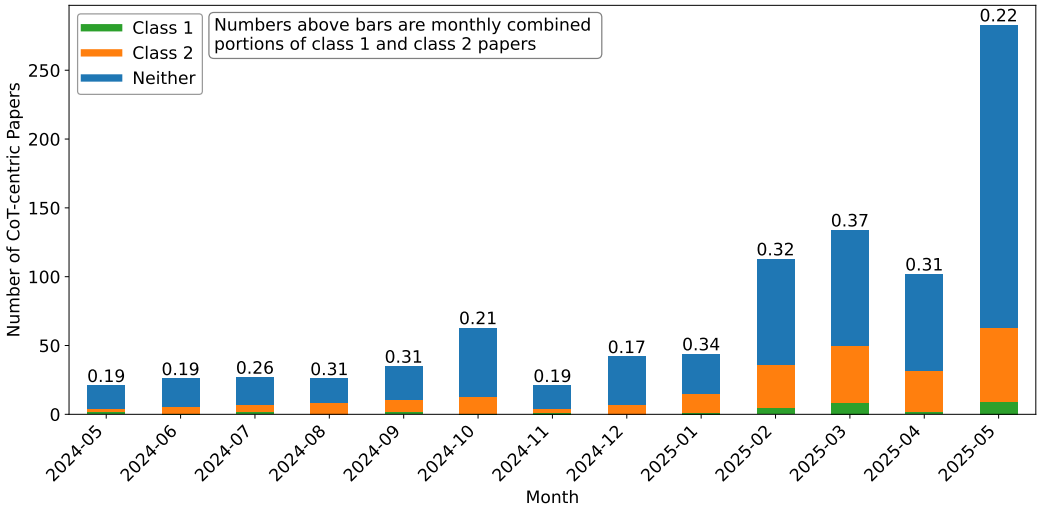
We outline our paper collection criteria and the query used for retrieving text chunks and constructing the final prompt. Specifically, we include an arXiv paper if:

1. Its abstract contains any of the following strings: "chain-based reasoning," "CoT," or "chain-of-thought."
2. It has a minimum length of 8 pages.

From the pool of papers meeting these criteria, we collect the 1,000 most recent papers.



(a) The class distribution. 24.4% of the papers describe CoT as a technique to enhance interpretability when incorporating CoT into their models or the construction of datasets.



(b) Monthly counts of the three paper classes, along with the combined proportion of class 1 and class 2 papers. Notably, the portion of CoT interpretability claim does not show a declining trend over time.

Figure 3: The statistics of the 1000 most recent CoT-centric papers collected from arXiv.

The input query is as follows:

Chain-of-thought is not interpretable/explainable/transparent because it may not reflect an LLM's internal computations. However, some papers still (1) mention chain-of-thought (or chain-based reasoning) as an interpretable/explainable/transparent technique; or even claim (2) they adopt chain-of-thought to establish an interpretable model/framework/pipeline/agent/dataset.

Does this paper claim (1)? Or even (2)? Or none of them (N)? Give me the answer with reasons following the template "answer: X reasons: ", where X is 1, 2, or N.

B.2 Results

Figure 3a presents the distribution of the three classes among the 1000 CoT-centric papers. We find that 24.4%—a non-negligible percentage—of the papers, when introducing their CoT-based

frameworks, regard CoT as a technique that has made their models interpretable in addition to performance gain. Only 3.4% of the papers do not link their core methods to interpretable CoT and only admit CoT as an interpretability technique. To explore temporal trends, we group papers by final update month and plot the combined portion of class 1 and class 2 papers, as shown in Figure 3b (2024-04 and 2025-06 are excluded due to incomplete data coverage). We observe no clear decline in interpretability claim, highlighting the motivation behind our work.

To assess the reliability of our automated pipeline, we manually classify the most recent 100 CoT-centric papers (from 2025-05-25 to 2025-06-05). The resulting agreement rate is 83%, with a false positive rate—cases where we label a paper as "neither" but GPT-4o labels it as "class 1" or "class 2"—of only 5%.

Finally, we adopt another simple GPT-4o-based classification pipeline to categorize each of the 1000 papers into one of the four domain classes—medical AI, AI for law, autonomous vehicles, and none-of-the-above—based on paper title and abstract. This automated classification is followed by manual verification to ensure 100% precision. Our analysis reveals that papers in high-stakes domains are more likely to frame CoT as an interpretability tool compared to the overall average (25%). Specifically, 16 of 42 (38%) medical AI papers, 17 of 27 (63%) autonomous vehicle papers, and 1 of 4 (25%) AI-for-law papers adopt this interpretability framing.