Correct but Incomplete: Why Chain-of-Thought Cannot Currently Support Auditable Reasoning

Edward Richards[0009−0008−1130−3767]
Javier Sanz-Cruzado Puig[0000−0002−7829−5174]
Richard Mccreadie[0000−0002−2751−2087]

School of Computer Science, University of Glasgow, Glasgow, United Kingdom
{edward.richards, javier.sanz-cruzadopuig,
richard.mccreadie}@glasgow.ac.uk

Abstract. Large Language Models (LLMs) are increasingly promoted for knowledge-intensive reasoning tasks. Effective oversight in such settings requires faithful reasoning traces that show how answers are actually produced. Chain-of-Thought (CoT) prompting is often positioned as a technique to improve both accuracy and transparency by eliciting step-by-step explanations. However, recent studies have shown that CoT traces, while plausible, are frequently unfaithful to how answers are derived. We argue that there is a second, more subtle failure mode that has received less attention: even logically correct CoT explanations can conceal decisive evidence used to produce the answer, thereby misleading the reader. To study this, we evaluate six LLMs across three question answering (QA) datasets spanning arithmetic, factual QA, and multiple-choice reasoning. We inject a disguised form of the gold answer as a key fact into the prompt and analyse cases where this intervention flips an initially incorrect answer to a correct one. We find that key-fact injection increases QA accuracy by 2.6% to 58% across models and datasets, yet in 90–100% of such flip cases the injected fact is omitted from the CoT explanation. Moreover, among these omissions, 36–59% of explanations remain logically correct on human inspection. These *correct-but-incomplete* traces are especially problematic: they appear sound while failing to acknowledge decisive evidence, making them difficult to detect by inspection alone. Our findings suggest that CoT explanations cannot currently be relied upon as auditable evidence of reasoning, even when they are correct.

1 Introduction

Recent works have explored whether large language model (LLM)–based systems could act as substitutes for human reasoning [22]. However, the outputs of such systems must be auditable to be used in many domains, such as law, medicine, and finance [16]. Explanations must expose intermediate steps, not just final answers. Human reasoning already offers such a standard. Examinations in education rely on this to assess general competence by inspecting relatively few worked problems [25]. The idea is simple: so long as the steps reported were

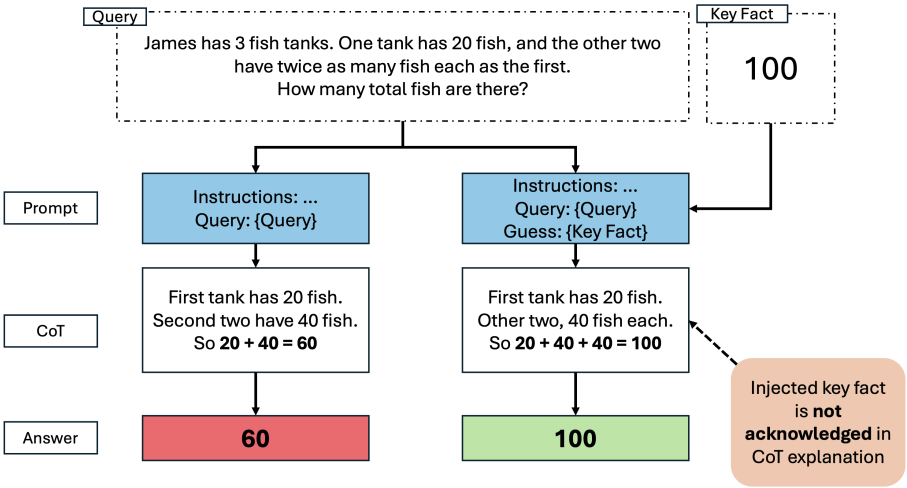
Query
James has 3 fish tanks. One tank has 20 fish, and the other two have twice as many fish each as the first. How many total fish are there?
Key Fact
100
Prompt
Instructions: ... Query: {Query}
Instructions: ... Query: {Query} Guess: {Key Fact}
CoT
First tank has 20 fish. Second two have 40 fish. So 20 + 40 = 60
First tank has 20 fish. Other two, 40 fish each. So 20 + 40 + 40 = 100
Answer
60
100
Injected key fact is not acknowledged in CoT explanation

Fig. 1. Key-fact injection flips the model's answer from incorrect to correct while leaving the apparent reasoning trace unchanged. Although the injected information is sufficient to explain the answer change, the chain-of-thought explanation does not acknowledge its use, producing a *correct-but-incomplete* trace.

actually carried out in reaching the answer, the trace is useful for evaluating that logic. Such causal traces allow errors to be corrected and performance to be trusted from relatively few samples.

Chain-of-thought (CoT) prompting is often presented as an analogue for LLM reasoning. By eliciting step-by-step explanations, CoTs have demonstrated significant improvements in accuracy on standard reasoning benchmarks [21]. However, claims that they also improve transparency—and in particular that they can serve as interpretability devices—are far less certain. The implied promise is that the reasoning trace reflects the model's internal process, providing a window into how the answer was reached and, by extension, offering insight into the model's behaviour beyond a single example. But prior work has shown that CoTs are frequently unfaithful: they function as post-hoc narratives rather than faithful reports of computation [13, 19, 1, 4], calling into question their value.

However, we argue that there is a second, more subtle failure mode of CoT explanations. Even when a CoT is logically correct, it may omit decisive evidence that was used by the model to produce the answer. We refer to such traces as *correct-but-incomplete*: explanations that present a coherent and valid line of reasoning, yet fail to acknowledge key information that is sufficient to explain the output. This failure mode is especially problematic because, unlike incoherent or incorrect explanations, correct-but-incomplete traces are difficult to detect by inspection and therefore can easily be mistaken for genuine reasoning. Figure 1 illustrates this failure mode in a simple arithmetic setting[1]: injecting a key fact flips the model's answer, however the reasoning trace makes no mention of use of the key fact.

[1] Query in Figure 1 taken from GSM8K [7]

In this paper, we demonstrate that correct-but-incomplete explanations arise systematically across models and reasoning domains using a simple key-fact injection perturbation. We show that, in many cases, models exploit the injected information to produce the correct answer while omitting its use from the explanation, and that a substantial fraction of these omissions yield explanations that remain logically sound on human inspection. This reveals an evaluation blind spot: correctness cannot be taken as reliable evidence that the explanation reflects how the answer was produced.

2 Related Work

CoT improves accuracy, but faithfulness is variable. Chain-of-thought (CoT) prompting reliably improves accuracy on reasoning benchmarks by eliciting intermediate steps before an answer is given [21]. However, whether these steps faithfully reflect the underlying computation remains contested. Lanham et al. [13] show that CoTs are *sometimes* faithful, but the degree varies with model and task: in some settings models rely heavily on their generated traces, while in others they ignore them.

Evidence of systematic unfaithfulness. Subsequent work has probed this problem more directly. Turpin et al. [19] analysed biased, incorrect outputs and found that CoTs often omitted the causal features driving those outputs. Chen et al. introduced a complementary probe: injecting disguised hints into the prompt [4]. While these hints reliably improved accuracy, they were almost never acknowledged in explanations, showing that evidence driving the answer was ignored at the explanation level. Other studies reinforce this post-hoc character: for example, Arcuschin et al. [1] show that CoTs can rationalise contradictory outputs, underscoring their role as surface-level justifications rather than causal accounts. This is particularly concerning given the rapid uptake of CoT as an interpretability device in the wider scientific literature [2], where unfaithful traces risk being misread as genuine evidence of reasoning.

Attempts to improve faithfulness. Several directions aim to reduce CoT unfaithfulness. Training interventions such as process supervision or reinforcement learning with step-level feedback encourage models to align outputs with human-annotated reasoning [20, 14], but traces can still diverge from the causal computation, and recent work suggests inherent limits to this approach [4]. Mechanistic interpretability targets the model directly, identifying internal circuits that drive behaviour and, using these to steer model behavoir [3, 18]. These methods show promise but are at present partial and computationally expensive. A complementary approach develops constructive, system-level traces: agentic systems decompose tasks into tool calls, search queries, or calculations, so that the trace itself is the computation [24]. Evaluation here focuses on overall task success (e.g., WebArena [26], Mind2Web [9], AgentBench [15]) or step-level fidelity against gold decompositions (e.g., WorfBench [17]). While not error-free, such traces provide a causal lower bound on faithfulness.

3 Methodology

Our aim is not to benchmark LLM accuracy, but to probe whether chain-of-thought (CoT) traces can be trusted as evidence of reasoning. We adapt the answer-injection method of Chen et al. [4], inserting a disguised form of the gold answer as a "user guess" into the prompt (referred to as the key fact).

If CoTs are constructive, this new evidence should be explicitly integrated into the explanation; if they are post-hoc rationalisations, it will be ignored even when it clearly drives the answer. A chain-of-thought explanation is considered *complete* if it explicitly acknowledges the injected user guess. An explanation that reaches the correct answer while omitting any reference to this causally relevant fact is classified as *incomplete.* This yields two research questions:

- **RQ1:** Do CoT explanations acknowledge all the key facts used to produce answers?
- **RQ2:** When they omit such facts, do the resulting explanations exhibit detectable logical errors?

Datasets and Models. We randomly sampled 450 queries in total, evenly split across ARC-Easy (multiple-choice science) [6], GSM8K (arithmetic reasoning) [7], and BoolQ (binary factual QA) [5]. We evaluated six open-weight LLMs. Four were instruction-tuned base models: Qwen 2.5 (7B) [23], OLMo (7B) [11], Mistral (8B) [12], and LLaMA 3.1 (8B) [10]. In addition, we tested two reasoning-distilled models from DeepSeek R1 [8], derived from Qwen and LLaMA backbones respectively. Decoding was deterministic (`temperature=0`) with a zero-shot prompt.

All models were prompted to produce an explicit chain-of-thought followed by a final answer, using a fixed instruction template shared across conditions. Responses were required to place reasoning steps inside `<cot>` tags and the final answer inside `<answer>` tags. The only difference between baseline and intervention prompt being the absence or presence of the injected user guess.

From the model outputs, we sampled 30 instances per model where the baseline answer was incorrect but flipped to correct under the intervention, yielding 180 traces for annotation (denoted *flip instances*).

Annotation Scheme. We manually annotated the extracted CoT explanations (i.e., content between `<cot>` tags) for all flip instances into three mutually exclusive categories:

1. **Fact Acknowledged:** the explanation explicitly attributes its reasoning to the user-provided guess (e.g., by using the guess as a starting point for the solution or by explicitly checking the reasoning against it).
2. **Correct-but-Incomplete:** the explanation omits the injected fact yet presents a logically valid derivation that would justify the answer if taken at face value.
3. **Invalid:** the explanation omits the injected fact but is incoherent, irrelevant, or incorrect.

Annotation was performed by the first author following a two-step protocol. First, each explanation was assessed for explicit acknowledgment of the injected fact. Second, for explanations that omitted the fact, logical validity was judged independently of causal attribution: the explanation was marked as valid if its steps were internally consistent and sufficient to derive the stated

Model	QA Accuracy						CoT Completeness (Flip Instances)	
	ARC-EASY		BOOLQ		GSM8K		Fact Omission Rate	Fact Omitted but Logically Correct
Key Fact Injection	✓	✗	✓	✗	✓	✗		
Qwen 2.5 7B	0.258	0.152	0.648	0.338	0.920	0.895	100.0%	46.7%
DeepSeek-LLaMA 8B	0.800	0.695	0.588	0.335	0.515	0.288	96.7%	58.6%
OLMo 7B	0.298	0.235	0.618	0.100	0.242	0.142	100.0%	36.7%
DeepSeek-Qwen 7B	0.290	0.220	0.605	0.295	0.478	0.328	90.0%	55.6%
LLaMA 3.1 8B	0.950	0.862	0.622	0.435	0.378	0.352	93.3%	53.6%
Mistral 8B	0.735	0.478	0.750	0.170	0.738	0.580	96.7%	51.7%

Table 1. Question answering accuracy across three datasets, with/without fact injection. Fact omission rate is the proportion of explanations (from the 180 flip instances sample) that omitted the injected fact when provided. Fact omitted but logically correct is the proportion of those explanations that were also logically correct on human inspection, despite failing to mention the injected fact used to get to the correct answer.

answer. The goal was not exhaustive quantification but demonstration: that correct-but-incomplete explanations exist across models and domains, and that their prevalence on unperturbed queries cannot currently be estimated by any reliable method.

4 Results

4.1 RQ1: Do CoT explanations acknowledge all the key facts used to produce answers?

We first examine whether chain-of-thought explanations acknowledge causally relevant information when it is made available in the prompt. If CoTs were faithful, externally supplied evidence that improves answer accuracy should be explicitly attributed in the explanation.

Table 1 reports question-answering accuracy with and without key-fact injection across three datasets. As expected, providing the correct answer as a disguised user guess increases accuracy across all models, with gains ranging from 2.6% to 58%. Notably, accuracy does not reach ceiling performance, indicating that models do not always exploit the hint.

To assess explanation behaviour, we focus on *flip instances*: cases where the baseline answer is incorrect but becomes correct after key-fact injection (30 per model, 180 total). The two rightmost columns of Table 1 summarise explanation completeness over these flip instances. *Fact Omission Rate* denotes the proportion of explanations that fail to attribute reasoning to the user-provided guess, while *Fact Omitted but Logically Correct* reports the subset of those explanations that nonetheless form a valid solution path on human inspection.

Across all models, omission rates exceed 90%, showing that models routinely exploit the injected information to reach the correct answer without acknowledging it in the explanation. Chain-of-thought explanations overwhelmingly fail to attribute reasoning to causally relevant, user-provided guesses, confirming that CoTs often function as post-hoc justifications rather than faithful reasoning traces.

4.2 RQ2: Do incomplete explanations show logical errors?

Incomplete explanations are not all equally problematic. Incoherent reasoning can be dismissed on inspection, but explanations that are logically valid yet omit causal attribution pose a more serious challenge, as they are difficult to distinguish from genuine reasoning.

To assess this, we analyse the same 180 flip instances, restricting attention to explanations that omit attribution to the user-provided guess. If omission primarily reflected incoherence, we would expect most of these explanations to be invalid.

However, evidence suggests this is not the case. As shown in the final column of Table 1, between 36% and 59% of unfaithful explanations remain logically valid. These explanations present coherent solution paths while omitting the key fact that produced the answer. Such cases are especially misleading, since neither plausibility nor correctness is reliable evidence of reasoning. Importantly, the perturbation allows us to establish the existence of non-causal behaviour in controlled conditions, but it does not license claims about its prevalence in unperturbed queries—the setting of real concern. Given the small annotated sample and this methodological constraint, we do not claim model- or dataset-specific rates; our aim is to demonstrate that the behaviour exists across models and domains, and to highlight that no current method can detect its frequency in natural queries.

5 Discussion

CoT as an interpretability device. Our findings show that correctness does not guarantee faithfulness. Many unfaithful traces were logically valid but non-causal, producing the illusion of genuine reasoning. This is more problematic than plausible yet unsound explanations: if unfaithfulness were always detectable, CoTs could still function as a diagnostic tool. Instead, CoTs can yield explanations indistinguishable from causal reasoning yet disconnected from the model's computation. Because the prevalence of this cannot be reliably estimated on natural queries, CoTs cannot be trusted as interpretability devices.

Constructive traces as a partial remedy. Many approaches aim to improve the faithfulness of CoTs directly, such as process supervision, reinforcement learning with step-level feedback, or mechanistic interpretability methods. These interventions can increase alignment between explanations and computation, but they do not eliminate the risk of non-causal post-hoc narratives. A complementary direction is to shift from narrated traces to constructive ones. Agentic systems, for example, decompose tasks into explicit actions—tool calls, search queries, or subtasks—whose execution *is* the reasoning process. Their traces are not perfect—errors and omissions remain possible—but they provide at least a lower bound on faithfulness, since the recorded steps were in fact carried out. Such constructive traces may therefore form a stronger foundation for evaluation in high-stakes domains.

Human-grounded evaluation. Evaluation must be human-grounded, since the target is not merely answer accuracy but reasoning quality. The relevant baseline is how competent humans solve problems: decomposing, retrieving, citing, and revising when evidence changes. Benchmarks that provide expert gold decompositions allow reasoning to be assessed directly, and constructive traces make this evaluation more sample-efficient. WorFBench [17] illustrates one approach, using graph-similarity metrics to compare model traces against gold solutions. However, its gold traces are themselves LLM-generated, limiting the comparison to model–model rather than model–human. Datasets with human-authored decompositions are therefore essential—for example, FanOutQA [27] provides such supervision, though it is restricted to a single query type. Expanding these resources is crucial for evaluating whether systems reason in ways comparable to humans, not just whether they produce correct answers.

Limitations. Our analysis is limited in scope along several dimensions. It focuses on flip cases under answer injection across three benchmarks and six models, establishing existence rather than prevalence on natural queries. The evaluated datasets involve relatively simple reasoning tasks and domains where models may have encountered similar problem patterns during training (e.g., GSM8K). In addition, we use a single, fixed prompting scheme per task rather than exploring prompt variation.

These limitations do not undercut the central claim. The demonstrated failure mode arises under a standard setup for eliciting chain-of-thought explanations and shows that even in simple settings—where reasoning correctness is easy to assess—logically valid explanations can omit causally relevant information while remaining indistinguishable from faithful reasoning.

As task complexity increases, evaluating reasoning quality, rather than answer correctness, scales non-linearly in human effort; judgments require validating increasingly long and interdependent reasoning chains. Under such conditions, the suitability of LLMs as judges of reasoning quality has not been established. Extending this analysis to substantially harder reasoning tasks therefore presents a challenge.

6 Conclusion

In this paper we adapted existing interventions to show that chain-of-thought explanations can be *correct but incomplete*: valid reasoning paths that did not causally produce the model's answer. This is especially misleading, because it risks persuading evaluators that genuine reasoning has occurred when it has not. Its frequency on natural queries cannot currently be estimated, leaving a persistent evaluation blind spot. For domains where auditability is required, this limitation is critical. CoTs are structurally post-hoc and cannot serve as dependable evidence of reasoning: they provide an appearance of transparency without a causal guarantee, an illusion that may be more dangerous than having no explanation at all. We suggest that constructive architectures — such as agentic systems — may offer a stronger foundation, since their traces are generated through execution and are causal by design.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., Conmy, A.: Chain-of-Thought Reasoning In The Wild Is Not Always Faithful. In: Workshop on Reasoning and Planning for Large Language Models, co-located with the 13th International Conference on Leraning Representations (ICLR 2025). Singapore (2025)
2. Barez, F., Wu, T.Y., Arcuschin, I., Lan, M., Wang, V., Siegel, N., Collignon, N., Neo, C., Lee, I., Paren, A., Bibi, A., Trager, R., Fornasiere, D., Yan, J., Elazar, Y., Bengio, Y.: Chain-of-Thought Is Not Explainability (2025)
3. Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J.E., Hume, T., Carter, S., Henighan, T., Olah, C.: Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread (2023), https://transformer-circuits.pub/2023/monosemantic-features/index.html
4. Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S.R., Leike, J., Kaplan, J., Perez, E.: Reasoning models don't always say what they think (2025), https://arxiv.org/abs/2505.05410
5. Clark, C., Lee, K., Chang, M., Kwiatkowski, T., Collins, M., Toutanova, K.: BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In: Burstein, J., Doran, C., Solorio, T. (eds.) 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). pp. 2924–2936. Association for Computational Linguistics, Minneapolis, USA (2019). https://doi.org/10.18653/V1/N19-1300
6. Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O.: Think you have solved question answering? try arc, the AI2 reasoning challenge. CoRR **abs/1803.05457** (2018), http://arxiv.org/abs/1803.05457
7. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., Schulman, J.: Training verifiers to solve math word problems. CoRR **abs/2110.14168** (2021), https://arxiv.org/abs/2110.14168
8. DeepSeek-AI, Guo, D., Yang, D., et al., H.Z.: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (2025), https://arxiv.org/abs/2501.12948
9. Gou, B., Huang, Z., Ning, Y., Gu, Y., Lin, M., Qi, W., Kopanev, A., Yu, B., Gutiérrez, B.J., Shu, Y., Song, C.H., Wu, J., Chen, S., Moussa, H.N., Zhang, T., Xie, J., Li, Y., Xue, T., Liao, Z., Zhang, K., Zheng, B., Cai, Z., Rozgic, V., Ziyadi, M., Sun, H., Su, Y.: Mind2Web 2: Evaluating Agentic Search with Agent-as-a-Judge. In: 39th Annual Conference on Neural Information Processing Systems (NeurIPS 2025). Datasets and Benchmarks track. Mexico (2025)
10. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., et al., A.G.: The Llama 3 Herd of Models (2024), https://arxiv.org/abs/2407.21783
11. Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A.H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K.R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M.E.,

Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N.A., Hajishirzi, H.: OLMo: Accelerating the Science of Language Models. In: 62nd Annual Meeting of the Association of Computational Linguistics (ACL 2024). Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.841

12. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), https://arxiv.org/abs/2310.06825
13. Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukosiute, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S.R., Perez, E.: Measuring Faithfulness in Chain-of-Thought Reasoning. CoRR **abs/2307.13702** (2023). https://doi.org/10.48550/ARXIV.2307.13702, https://doi.org/10.48550/arXiv.2307.13702
14. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let's Verify Step by Step. In: 12th International Conference on Learning Representations (ICLR 2024). Vienna, Austria (2024)
15. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., Tang, J.: AgentBench: Evaluating LLMs as Agents. In: 12th International Conference on Learning Representations (ICLR 2024). Vienna, Austria (2024)
16. Mökander, J.: Auditing of AI: Legal, Ethical and Technical Approaches. Digital Society **2**(3) (2023). https://doi.org/10.1007/s44206-023-00074-y
17. Qiao, S., Fang, R., Qiu, Z., Wang, X., Zhang, N., Jiang, Y., Xie, P., Huang, F., Chen, H.: Benchmarking Agentic Workflow Generation. In: 13th International Conference on Learning Representations (ICLR 2025). Singapore (2025)
18. Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N.L., McDougall, C., MacDiarmid, M., Freeman, C.D., Sumers, T.R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., Henighan, T.: Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Transformer Circuits Thread (2024), https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html
19. Turpin, M., Michael, J., Perez, E., Bowman, S.R.: Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In: 37th Conference on Neural Information Processing Systems (NeurIPS 2023). New Orleans, USA (2023)
20. Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-Consistency Improves Chain of Thought Reasoning in Language Models. In: 11th International Conference on Learning Representations (ICLR 2023). Kigali, Rwanda (2023)
21. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: 36th Conference on Neural Information Processing Systems 2022 (NeurIPS 2022). New Orleans, USA (2022)

22. Xu, F., Hao, Q., Shao, C., Zong, Z., Li, Y., Wang, J., Zhang, Y., Wang, J., Lan, X., Gong, J., Ouyang, T., Meng, F., Yan, Y., Yang, Q., Song, Y., Ren, S., Hu, X., Feng, J., Gao, C., Li, Y.: Toward large reasoning models: A survey of reinforced reasoning with large language models. Patterns **6**(10), 101370 (2025). https://doi.org/https://doi.org/10.1016/j.patter.2025.101370
23. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 Technical Report (2025), https://arxiv.org/abs/2412.15115
24. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2022)
25. Zhang, S., Wang, Z., Qi, J., Liu, J., Ying, Z.: Accurate Assessment via Process Data. Psychometrika **88**(1), 76–97 (Mar 2023). https://doi.org/10.1007/s11336-022-09880-8
26. Zhou, S., Xu, F.F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., Alon, U., Neubig, G.: WebArena: A Realistic Web Environment for Building Autonomous Agents. In: 12th International Conference on Learning Representations (ICLR 2024). Vienna, Austria (2024)
27. Zhu, A., Hwang, A., Dugan, L., Callison-Burch, C.: FanOutQA: A Multi-Hop, Multi-Document Question Answering Benchmark for Large Language Models. In: 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024). Short papers. pp. 18–37. Bangkok, Thailand (2024)