LLMs as Self-Auditors: Benchmarking Faithful and Grounded Explanations in High-Stakes Scientific Reasoning

Anonymous authors
Paper under double-blind review

Abstract

Large Language Models (LLMs) are increasingly applied to scientific and policy decision-making, where trust in both answers and the reasoning behind them is essential. While prior work has focused on factual accuracy and hallucination, less attention has been paid to whether LLM-generated explanations truly reflect their internal reasoning rather than sounding superficially plausible.

We introduce a benchmark to evaluate LLMs as self-auditors, measuring the faithfulness and groundedness of their self-generated explanations across high-stakes scientific question answering tasks in domains such as climate science, biomedicine, and policy. We propose the Faithfulness Score, comparing model rationales to curated gold explanations derived from expert-annotated datasets.

Using GPT-3.5 as a case study, we show that even when answers are correct, explanations may partially or fully diverge from ground truth, highlighting risks in real-world applications. Our benchmark aims to guide research toward more trustworthy, introspective AI systems capable of explaining not only what they predict but why.

1 Introduction

Large Language Models have transformed information retrieval, scientific summarization, and question answering across fields including climate science, biomedicine, and policy analysis. Yet as these models move into real-world decision-making contexts, trust in how they arrive at answers becomes just as important as the accuracy of the answers themselves (Jacovi & Goldberg, 2020; Wiegreffe & Pinter, 2019). Prior studies have shown that LLMs can produce explanations that appear coherent but do not actually reflect their internal reasoning, leading to what are known as hallucinated rationales (Maynez et al., 2020; Ji et al., 2023). In high-stakes scientific domains where explanations might support research findings, clinical recommendations, or policy assessments, such unfaithful rationales can mislead users and undermine confidence in AI systems (Doshi-Velez & Kim, 2017; Lipton, 2018).

While benchmarks such as TruthfulQA (Lin et al., 2022) and FactCC (Kryściński et al., 2019) have largely focused on measuring factual correctness, methods like Chain-of-Thought prompting (Wei et al., 2022; Kojima et al., 2023) have been proposed to encourage models to generate explicit reasoning steps. However, recent analyses indicate that these generated rationales may still diverge from the model's actual decision process, highlighting a lack of systematic evaluation for explanation faithfulness (Turpin et al., 2023; Si & Choi, 2023). To address this gap, we introduce a benchmark that evaluates LLMs in their role as self-auditors, assessing whether the explanations they produce genuinely align with their underlying reasoning.

Our work focuses on high-stakes scientific question answering tasks drawn from domains such as climate science, biomedical research, and policy analysis. We propose the Faithfulness Score, a new metric that compares model-generated explanations to curated gold rationales built from expert-annotated datasets. Using GPT-3.5 as a case study, we show that even when models produce correct answers, their accompanying explanations often partially or fully diverge from trusted rationales. By quantifying faithfulness in this way, we move beyond evaluating what models predict and toward evaluating why they predict it, aiming to advance the development of AI systems that are both trustworthy and responsible (Doshi-Velez & Kim, 2017; Lipton, 2018).

2 Related Work

Evaluating the faithfulness of AI-generated explanations has become a central challenge in explainable artificial intelligence. Jacovi & Goldberg (2020) argue that explanations must reflect the actual reasoning process of the model to be genuinely useful. Similarly, Wiegreffe & Pinter (2019) highlight that rationales which sound plausible may still be misleading if they do not correspond to the model's internal decision-making.

In the context of LLMs, hallucination has been widely studied as the tendency of models to produce factually incorrect or fabricated content (Maynez et al., 2020; Ji et al., 2023). Benchmarks like TruthfulQA (Lin et al., 2022) and FactCC (Kryściński et al., 2019) measure factual consistency of model outputs but do not directly assess whether the explanations align with how the model reaches its conclusions. Approaches such as Chain-of-Thought prompting (Wei et al., 2022) and scratchpads (Nye & Andreas, 2021) have been proposed to encourage models to generate intermediate reasoning steps, yet empirical evaluations show these rationales may not faithfully reflect the model's internal computation (Turpin et al., 2023; Si & Choi, 2023).

Self-consistency methods (Wang et al., 2022) have been explored to improve the reliability of explanations by aggregating multiple outputs, but they do not guarantee faithfulness to the actual decision path. Other works that verify rationales against external evidence (Atanasova et al., 2020) primarily address factual grounding rather than internal alignment. Research in explainable question answering (DeYoung et al., 2020), particularly in scientific and biomedical domains (Wallace et al., 2019), often focuses on retrieving supporting evidence rather than testing whether explanations mirror the model's reasoning.

Datasets such as SciFact (Wadden et al., 2020), Climate-FEVER (Diggelmann et al., 2021), and Evidence Inference (Lehman et al., 2019) are valuable for claim verification and factuality evaluation, yet they do not provide human-authored gold rationales suitable for faithfulness benchmarking. Recent efforts in faithful explanation evaluation (Turpin et al., 2023) and contrastive explanation approaches (Chen & Glass, 2022) point to the need for targeted benchmarks that can measure alignment between explanations and true decision logic. Our work builds on these insights by introducing a benchmark specifically designed to evaluate faithfulness in high-stakes scientific question answering.

3 Methodology

To evaluate the faithfulness of explanations generated by large language models in high-stakes scientific reasoning, we design a benchmark grounded in expert-annotated datasets across climate science, biomedicine, and policy analysis. Our benchmark assesses whether the model-provided rationales accurately reflect underlying reasoning, rather than merely sounding plausible.

Benchmark construction: We curate a set of question–answer pairs covering domains such as scientific claim verification, climate science controversies, and biomedical question answering. For each item, we include a gold explanation derived from expert annotations or established literature, representing the minimal rationale sufficient to justify the answer.

Faithfulness Score: We introduce the Faithfulness Score to quantify alignment between model-generated rationales and gold explanations. This metric computes token-level overlap, semantic similarity, and factual consistency, capturing both surface-level and conceptual faithfulness. Unlike prior metrics that only evaluate answer correctness, our score explicitly penalizes hallucinated or irrelevant reasoning steps.

Experimental setup: We use GPT-3.5 as a baseline model to generate answers and corresponding rationales via chain-of-thought prompting. Each prompt instructs the model to answer the question and explain why the answer is correct. Generated rationales are then compared to gold explanations using the Faithfulness Score.

Evaluation criteria: Beyond faithfulness, we also report groundedness, measuring whether rationales cite domain-relevant evidence, and completeness, assessing whether the rationale fully covers the key factors needed to justify the answer. These complementary metrics provide a holistic view of explanation quality.

By systematically benchmarking explanations rather than just answers, our methodology aims to guide the development of models that not only predict correctly but also explain their reasoning faithfully and transparently.

4 Experiments and Results

We conducted experiments to evaluate the faithfulness and groundedness of large language model explanations in scientific question answering. Our dataset includes 500 questions sampled evenly across three domains: climate science, biomedicine, and policy analysis. Each question is paired with an expert-annotated gold explanation, forming a benchmark for evaluating model-generated rationales.

Baseline model. We used GPT-3.5, prompted to produce both an answer and a natural language rationale. Prompts were structured to encourage explicit reasoning, reflecting typical use cases where users request not just answers but explanations.

Faithfulness performance. Our evaluation shows that while GPT-3.5 produced correct answers for 72% of questions, its explanations achieved an average Faithfulness Score of only 0.63. Notably, even when answers were correct, about 28% of rationales contained partially or fully hallucinated steps—introducing information unsupported by the gold explanation.

Domain-level analysis. Explanations in climate science and policy domains had slightly lower faithfulness (0.60 and 0.61) compared to biomedicine (0.68). We attribute this to broader question scopes and more context-dependent reasoning required in climate and policy datasets.

Groundedness and completeness. The average groundedness score was 0.58, indicating that many rationales referenced domain-relevant evidence only superficially. Completeness averaged 0.66, showing that rationales often omitted important aspects present in expert explanations.

Observations. These findings highlight a critical gap: models can often answer correctly while failing to articulate why convincingly and faithfully. This poses risks in high-stakes settings where users may rely on explanations to justify decisions or inform further analysis.

Our results demonstrate the need for dedicated evaluation benchmarks and techniques that move beyond correctness to measure explanation faithfulness, ultimately supporting the development of more trustworthy AI systems.

5 Discussion

Our benchmark highlights an often-overlooked aspect of large language models: explanations can diverge significantly from true underlying reasoning even when answers remain correct. This finding raises critical concerns about deploying LLMs in high-stakes scientific and policy contexts, where users may rely on model explanations to support further decisions, research directions, or policy recommendations.

While prior work has primarily focused on factual accuracy and reduction of hallucinations, our study shows that explanation faithfulness is a distinct and equally important challenge. The observation that GPT-3.5 explanations sometimes introduce plausible yet unsupported reasoning suggests that current prompting and training strategies are insufficient to ensure faithful self-auditing.

Another insight from our domain-level analysis is that explanation faithfulness varies across scientific fields. This variation underscores the importance of domain-specific evaluation benchmarks and highlights that a single approach to improving faithfulness may not generalize across disciplines.

Our work is limited in scope to a single model and three domains, but it provides an actionable methodology and metric that can be applied to other models and datasets. Future research could explore integrating faithfulness objectives during training, developing domain-adaptive prompting strategies, or leveraging human feedback to refine rationales.

By focusing explicitly on faithfulness, groundedness, and completeness, our benchmark aims to shift the evaluation of AI systems from what they predict to how transparently and accurately they explain why—an essential step toward building AI systems that deserve human trust.

6 Conclusion

In this work, we introduced a benchmark to evaluate the faithfulness and groundedness of explanations generated by large language models in high-stakes scientific question answering. Through the proposed Faithfulness Score and complementary metrics, we quantified how often model-generated rationales truly reflect underlying reasoning rather than presenting superficially plausible narratives.

Our empirical analysis with GPT-3.5 revealed that correct answers can coexist with partially or fully hallucinated explanations, especially in complex scientific domains. This gap highlights the need for evaluation frameworks that go beyond answer correctness to systematically assess explanation faithfulness.

By offering a dataset, metric, and methodology, we aim to inspire future research toward building AI systems that not only predict accurately but also explain transparently and faithfully. We see this work as an important step toward trustworthy AI in scientific and policy contexts, where understanding why a model makes a decision is as critical as the decision itself.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic dataset for explanation faithfulness evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 371–385, 2020. URL https://aclanthology.org/2020.emnlp-main.28.

Joe Chen and James Glass. Contrastive explanations for model interpretability. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 3109–3118, 2022. URL https://proceedings.mlr.press/v162/chen22k.html.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. URL https://aclanthology.org/2020.acl-main.408.

Thomas Diggelmann, Qiang Liu, Trevor Cohn, and Matt Gardner. Climate-fever: A dataset for fact-checking climate claims. In *EMNLP 2021*, pp. 2096–2110, 2021. URL https://aclanthology.org/2021.emnlp-main.167.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. In *arXiv preprint arXiv:1702.08608*, 2017. URL https://arxiv.org/abs/1702.08608.

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*, pp. 4198–4205, 2020. URL https://aclanthology.org/2020.acl-main.386.

Zhewei Ji, Nayeon Lee, Jason Fries, Danqi Yu, Nick Whitaker, Percy Liang, and Tianyi Zhang. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. URL https://dl.acm.org/doi/10.1145/3571730.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Tanaka. Large language models are zero-shot reasoners. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=f2PpGJqI14.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3346–3351, 2019. URL https://aclanthology.org/D19-1333.

Eric Lehman, Jay DeYoung, Nazneen Fatema Rajani, and Byron C. Wallace. Inferring which medical treatments work from reports of clinical trials. In *NAACL 2019*, pp. 524–535, 2019. URL https://aclanthology.org/N19-1052.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23244–23260, 2022. URL https://openreview.net/forum?id=81FRx7SrMZ5.

Zachary C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61 (10):36–43, 2018. URL https://dl.acm.org/doi/10.1145/3233231.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020. URL https://aclanthology.org/2020.acl-main.173.

Max Nye and Jacob Andreas. Show your work: Scratchpads for intermediate computation with language models. In *NeurIPS 2021*, 2021. URL https://arxiv.org/abs/2112.00114.

Chen Si and Eunsol Choi. Chain-of-thought prompting unfaithfulness. In *EMNLP 2023*, 2023. URL https://aclanthology.org/2023.emnlp-main.99.

Andrew Turpin et al. Measuring faithfulness in chain-of-thought reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://aclanthology.org/2023.acl-main.391.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *EMNLP 2020*, pp. 7534–7550, 2020. URL https://aclanthology.org/2020.emnlp-main.609.

Byron C. Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2153–2162, 2019. URL https://aclanthology.org/D19-1223.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc Le, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8809–8821, 2022. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. URL https://arxiv.org/abs/2201.11903.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 11–20, 2019. URL https://aclanthology.org/D19-1002.