
Large Language Models Cannot Explain Themselves

Advait Sarkar

Microsoft Research
Cambridge, United Kingdom

University of Cambridge
Cambridge, United Kingdom

University College London
London, United Kingdom

advait@microsoft.com

Abstract

Large language models can be prompted to produce text. They can also be prompted to produce “explanations” of their output. But these are not really explanations, because they do not accurately reflect the mechanical process underlying the prediction. The illusion that they reflect the reasoning process can result in significant harms. These “explanations” can be valuable, but for promoting critical thinking rather than for understanding the model. I propose a recontextualisation of these “explanations”, using the term “explanations” to draw attention to their exogenous nature. I discuss some implications for design and technology, such as the inclusion of appropriate guardrails and responses when models are prompted to generate explanations.

Author Keywords

explanations; mechanisml explanations; co-audit; AI safety; explainable AI; XAI; critical thinking

CCS Concepts

•**Human-centered computing** → *HCI theory, concepts and models*; •**Natural language interfaces**; •**Computing methodologies** → Natural language processing; Neural networks; Machine learning; Philosophical/theoretical foundations of artificial intelligence;

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Copyright held by the owner/author(s).
CHI'24, Honolulu, HI, USA

Key terms

Mechanismal explanations:

explanations of AI model behaviour which represent facts about the underlying mechanisms of prediction, such as the model structure, training data, or model weights. They are generated from introspection of the model, its training data, and its inference process. Examples include LIME and SHAP.

Explanations: statements which appear to be explanations of AI output, but are not (and cannot be) a grounded reflection of the mechanism that generated that output. This is what language models produce when asked to “explain” themselves.

The illusion of explanation

In the context of Artificial Intelligence (AI), the term “explanation” can encompass many types of information. The most well-studied category of explanations is concerned with providing descriptions of the structure of a model, its training data, and most commonly, elaboration of any given output in terms of the algorithmic process followed to produce that output [29]. What these explanations have in common is that they aim to faithfully represent some aspect of the real underlying algorithmic mechanism of an AI model. Let us therefore refer to these as *mechanismal explanations*.¹

Classic examples of mechanismal explanations include LIME [26], SHAP [21], saliency maps [36, 33], and Kulesza et al.’s visualisations of Bayes classifiers [15], and Sarkar et al.’s visualisations of k-NN models [31]. Mechanismal explanations are not the only kind: researchers in recent years have carefully drawn attention to aspects of AI explanation that instead pertain to the socio-technical system in which AI is embedded [7, 6, 9].

A disconnection is now immediately visible between a classic mechanismal explanation, and what is produced when a language model is prompted to generate an explanation. The former is truly generated from, and has a concrete, grounded relation to, the actual processes and behaviour of a model. But a language model “explanation” has no such property. This has been previously noted [1, 19], but the reason for the problem is treated as self-evident. I would like to expand on these observations, to explain why so-called “self-explanations” are considered to be ungrounded.

¹The aim of this awkward construction is to emphasise that they are explanations *of a mechanism*; using the term “mechanical” or “mechanistic” might connote that the explanations themselves are generated mechanically, which is usually true but not relevant.

Let us examine what is actually happening when a language model has produced some output O, and is then prompted to give an explanation E for O. The process of generating E is simply another execution of the language model. E is a text composed through a sequence of next-token predictions, stochastically optimised to satisfy the query. E is not the result of an introspective reflection on the algorithmic process that was followed to generate O. A true mechanismal explanation would invoke, for example, some reference to the actual training data, model parameters, or activations, that were involved in the production of O. But if this meta-information about the prediction process is not accessible to the model to draw upon in generating E, it is theoretically impossible that E could reflect it, accurately or otherwise.

The situation is no different if E and O are requested in a single prompt, e.g., “*What is the capital of France? Explain your answer.*”, as opposed to two separate prompts or conversational turns, the first asking the question and the second asking for the explanation. In the single-turn case, and in multi-turn systems where previous responses are included in the query context, it is true that the generation of O is affected by the presence of an E-request in the query context, and vice versa. For example, the language model may well produce a more coherent and well-justified output if it can simultaneously attend to a fragment of language in which an explanation is requested. However, the notional E portion of the response still lacks a *mechanismal* grounding.

Statements of type E, then, are not explanations, at least not in the sense that the word is most commonly used in explainable AI research, and, we shall see, not even in the sense that users colloquially expect from these systems. They do not hold the epistemically privileged status over

statements of the type O that they claim or that people expect. In fact, they are outputs like any other. E-type statements could be described as justifications, or “post-hoc rationalisations” [25], but even these terms imply a greater degree of reflexivity and introspection than is warranted. They are simulacra of justification, or of rationalisation; samples from the space of texts with the shape of justifications.

Let us instead call them *explanations*. This term retains all the connotations of explanations (they may or may not be correct, they carry the appearance of insight, they often appeal to cause, logic, or authority), but explicitly captures the fact that they are *exogenous* to, outside of, the output they explain. They inhabit the same plane of reasoning as their object; they cannot look any further beneath the object than the object itself can.

Despite state-of-the-art performance on reasoning tasks [25, 35, 38], and one study that reported feature attribution explanations with performance comparable to LIME [13], recent work has delivered significant evidence that language models consistently fail to accurately explain their own output, and can even systematically misrepresent the true reason for a model’s prediction [4, 22, 32, 34]. In other words, at present, when the explanation sought requires introspection into the generation process, explanations just don’t work. Large language models cannot explain themselves.

This does not mean that explanations are not useful; on the contrary, when presented appropriately, they can be an important and powerful tool in the designer’s toolkit for creating useful and trustworthy experiences. Before we discuss those, let us turn our attention briefly to why it is important to make the distinction between explanations and explanations, beyond academic pedantry.

Societal harms of explanations

The story of the New York lawyers who submitted a legal brief including case citations generated by ChatGPT, but which turned out to be non-existent, is now well-known [23].

A less well-known aspect of this episode is that the infelicitous lawyers did attempt to verify that the cases were real... by asking ChatGPT, which confidently explained that the cases were real: *“[The lawyer] asked the AI tool whether [a case it generated] is a real case. ChatGPT answered that it “is a real case” and “can be found on legal research databases such as Westlaw and LexisNexis.” When asked if the other cases provided by ChatGPT are fake, it answered, “No, the other cases I provided are real and can be found in reputable legal databases such as LexisNexis and Westlaw.”*” [3]

It has often been noted that language model hallucinations are particularly dangerous because of the bold confidence with which the model makes its assertions. The same is true of explanations. Because it is so easy to prompt a language model to produce an explanation, which is reported with bold confidence, the user can be forgiven for thinking that explanations are mechanistic explanations, whereas in fact they are not. This can lead to the very obvious problems such as the example above. As the firm stated in response to the judgment that the lawyers had acted in bad faith, *“We made a good faith mistake in failing to believe that a piece of technology could be making up cases out of whole cloth”* [23].

As designers we must ask ourselves: in whom (or what) was this “good faith” placed, and why? If a false statement presented with bold confidence is dangerous, a false statement presented and explained with bold confidence is doubly so. Research in social psychology has shown that additional information can increase persuasiveness, even if

Harms of explanations

False confidence: bold explanations of hallucinated statements can give users false confidence in those statements, with dangerous consequences.

Diminished critical thinking: instead of engaging in introspection or evaluating the logic and evidence behind the model's output, users may accept explanations at face value.

Erosion of trust: when users discover that explanations do not accurately explain language model behaviour, this can undermine the credibility of AI systems.

it is irrelevant to the request [17]. Users are easily influenced and can place their trust in meaningless explanations [10], and can over-trust interpretability aids [14]. Allowing a system to present explanations with the veneer of explanations, in a situation where the user expects an explanation, should therefore be considered a dark pattern [2].

The illusion of explanation perpetuated by explanations poses a threat to decision-making processes, in everyday knowledge work as well as in high-stakes environments such as legal or medical contexts. Reliance on explanations may diminish users' critical thinking and decision-making abilities.

Instead of engaging in introspection or evaluating the logic and evidence behind the model's output, users may accept explanations at face value. And why shouldn't they? Computers are tools, and tools are not viewed as being adversarial to the activity they facilitate. It does not seem to be a productive avenue for interaction design to attempt to erase the cultural, inertial tendency to trust computers as computationally correct machines, even if that tendency is wildly misplaced in language models.

Explanations can also impair user trust and confidence in AI systems in the long term. As explanations are revealed to not, in fact, have their putative explanatory power, this can erode trust, and undermine any legitimate credibility that AI systems might have.

Recontextualising ex(o)planations

This is clearly a case that calls for a social construction of explainability, which should *"start with 'who' the relevant stakeholders are, their explainability needs, and justify how a particular conception of explainability satisfies the shared goals of the relevant social group"* [8].

It is not that mechanistic explanations for language models are lacking. Despite significant challenges [29], numerous techniques have been developed to explain feature attribution, neuron activation, model attention, etc. [37, 19].

However, mechanistic explanations are not the aim in and of themselves; the important aspect of the user experience that explanations need to fulfil is *decision support* [31, 28, 19, 11]. Is the output correct? If it isn't, what do I need to do to fix it? Can I trust this? For example, in an AI system that generates spreadsheet formulas from natural language queries, it is by far more important and consequential for the user experience to explain the generated *formula*, what it does and how it works, rather than the *mechanism* of the language model that produced it. Mechanistic explanations may generate confusion and information overload in such a context [16, 19].

As Miller [24] and Sarkar [29] have noted, human-human explanations are generally not mechanistic, in the sense that human-generated explanations of human behaviour rarely invoke low-level psychological or neurological phenomena, yet they are still generally successful at fulfilling the needs of everyday communication. Effective explanations can be contrastive, counterfactual, and justificatory, with respect to some intended state of affairs; these have nothing to do with the causal mechanisms underlying behaviour.

Parts of the decision support problem can be addressed though an approach termed "co-audit" [12]: tools to help check AI-generated content. An example of this would be the "grounded utterances" generated through a separate and deterministic mechanism to explain the model output [20]. Another technique, employed by Microsoft Copilot (formerly Bing Chat) is to cite references to its Web sources that can be followed and verified. These are true

explanations: they rely on mechanisms and authorities separate from the model itself and with an epistemically privileged view over the output generation process.

But explanations themselves can also be useful. Without needing to introspect the model, they can generate statements which help the *user* rationalise, justify, and evaluate. They can generate text that prompts the *user* to reflect on the output and their intents. Explanations can thus promote critical thinking about interactions with generative AI [30].

I propose a simple design implication that can be applied immediately: the introduction of guardrails and interface warnings against explanations. Commercial systems such as ChatGPT already abound with guardrails against content deemed inappropriate by the system designers, such as violent or sexual content, and numerous disclaimers against hallucinations, to the effect of *“AI generated content may be incorrect.”* To these considerations, I suggest adding guardrails against explanations masquerading as explanations, and contextualising them to allow their true and appropriate utility to shine.

For example, if the user asks the system to explain its output, it could produce a disclaimer of the following type: *“You asked for an explanation, but as a language model, I am incapable of explaining my own behaviour.”* It might then follow this with *“However, I can provide examples of how to justify, rationalise, or evaluate my previous response. Here are example arguments for and against it. This is not an explanation of my previous response.”* Together, such a disclaimer followed by an explanation could help defuse the worst dangers and infuse some critical thought.

There is reason to believe that such simple interventions can have a meaningful effect. The presence of

metacognitive guiding questions, such as “what do I understand from the text so far?” significantly improves reading comprehension [27]. Framing explanations as questions improves human logical discernment [5]. When technology sparks conflict in discussions, it improves critical thinking [18]. Users are influenced by the language of conversational systems and can change their instructional vocabulary and grammar after just a single exposure to system output [20]. The very same forces that influence and nudge users into trusting false explanations can be marshalled for their benefit instead.

Going forward, as more *true* explanation mechanisms are developed: co-audit tools, grounded utterances, citations, etc., such disclaimers may be replaced with more concrete decision-support mechanisms. However, the utility of explanations as critical thinking support will remain. The key will be in helping the user develop safe and effective behaviours and mental models of trust around the different sources of evaluation and reflection available.

Acknowledgements

Thanks to my reviewers for their time and feedback.

REFERENCES

- [1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
- [2] H Brignull, M Leiser, C Santos, and K Doshi. 2023. Deceptive patterns – user interfaces designed to trick you. (4 2023). <https://www.deceptive.design/>
- [3] Jon Brodtkin. 2023. Lawyer cited 6 fake cases made up by CHATGPT; judge calls it “unprecedented”. (Jun 2023). <https://arstechnica.com/tech-policy/2023/05/lawyer-cited-6-fake-cases-made-up-by-chatgpt-judge-calls-it-unprecedented/>
- [4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
- [5] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don’t Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [6] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021a. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [7] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, and others. 2021b. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509* (2021).
- [8] Upol Ehsan and Mark O. Riedl. 2022. Social Construction of XAI: Do We Need One Definition to Rule Them All? (2022).
- [9] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O Riedl. 2023. Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.
- [10] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA ’19)*. Association for Computing Machinery, New York, NY, USA, 1–6. DOI : <http://dx.doi.org/10.1145/3290607.3312787>
- [11] Raymond Fok and Daniel S. Weld. 2024. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. (2024).
- [12] Andrew D Gordon, Carina Negreanu, José Cambronero, Rasika Chakravarthy, Ian Drosos, Hao Fang, Bhaskar Mitra, Hannah Richardson, Advait Sarkar, Stephanie Simmons, and others. 2023. Co-audit: tools to help humans double-check AI-generated content. *arXiv preprint arXiv:2310.01297* (2023).

- [13] Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207* (2023).
- [14] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. DOI: <http://dx.doi.org/10.1145/3313831.3376219>
- [15] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.
- [16] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.
- [17] Ellen J Langer, Arthur Blank, and Benzion Chanowitz. 1978. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of personality and social psychology* 36, 6 (1978), 635.
- [18] Sunok Lee, Dasom Choi, Minha Lee, Jonghak Choi, and Sangsu Lee. 2023. Fostering Youth's Critical Thinking Competency About AI through Exhibition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 451, 22 pages. DOI: <http://dx.doi.org/10.1145/3544548.3581159>
- [19] Q. Vera Liao and Jennifer Wortman Vaughan. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review* (feb 29 2024). <https://hdsr.mitpress.mit.edu/pub/aelq19qy>.
- [20] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D Gordon. 2023. "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–31.
- [21] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
- [22] Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from Large Language Models faithful? (2024).
- [23] Sara Merken. 2023. New York lawyers sanctioned for using fake ChatGPT cases in legal brief. (Jun 2023). <https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/>
- [24] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

- [25] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361* (2019).
- [26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
- [27] Gavriel Salomon. 1988. AI in reverse: Computer tools that turn cognitive. *Journal of educational computing research* 4, 2 (1988), 123–139.
- [28] Advait Sarkar. 2016. *Interactive analytical modelling*. Technical Report UCAM-CL-TR-920. University of Cambridge, Computer Laboratory. DOI : <http://dx.doi.org/10.48456/tr-920>
- [29] Advait Sarkar. 2022. Is explainable AI a race against model complexity?. In *Workshop on Transparency and Explanations in Smart Systems (TeXSS), in conjunction with ACM Intelligent User Interfaces (IUI 2022) (CEUR Workshop Proceedings)*. 192–199. <http://ceur-ws.org/Vol-3124/paper22.pdf>
- [30] Advait Sarkar. 2024. AI Should Challenge, Not Obey. *Communications of the ACM (in press)* (2024).
- [31] Advait Sarkar, Mateja Jamnik, Alan F. Blackwell, and Martin Spott. 2015. Interactive visual machine learning in spreadsheets. In *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 159–163. DOI : <http://dx.doi.org/10.1109/VLHCC.2015.7357211>
- [32] Dane Sherburn, Bilal Chughtai, and Owain Evans. 2024. Language Models Struggle to Explain Themselves. (2024). <https://openreview.net/forum?id=o6eUNPBAEc>
- [33] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
- [34] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36 (2024).
- [35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [36] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I* 13. Springer, 818–833.
- [37] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology* 15, 2 (2024), 1–38.
- [38] Jiachen Zhao, Zonghai Yao, Zhichao Yang, and Hong Yu. 2023. SELF-EXPLAIN: Teaching Large Language Models to Reason Complex Questions by Themselves. *arXiv preprint arXiv:2311.06985* (2023).