

Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models

Chirag Agarwal¹ Sree Harsha Tanneru¹ Himabindu Lakkaraju¹

Abstract

Large Language Models (LLMs) are deployed as powerful tools for several natural language processing (NLP) applications. Recent works show that modern LLMs can generate self-explanations (SEs), which elicit their intermediate reasoning steps for explaining their behavior. Self-explanations have seen widespread adoption owing to their conversational and plausible nature. However, there is little to no understanding of their faithfulness. In this work, we discuss the dichotomy between faithfulness and plausibility in SEs generated by LLMs. We argue that while LLMs are adept at generating plausible explanations – seemingly logical and coherent to human users – these explanations do not necessarily align with the reasoning processes of the LLMs, raising concerns about their faithfulness. We highlight that the current trend towards increasing the plausibility of explanations, primarily driven by the demand for user-friendly interfaces, may come at the cost of diminishing their faithfulness. We assert that the faithfulness of explanations is critical in LLMs employed for high-stakes decision-making. Moreover, we urge the community to identify the faithfulness requirements of real-world applications and ensure explanations meet those needs. Finally, we propose some directions for future work, emphasizing the need for novel methodologies and frameworks that can enhance the faithfulness of self-explanations without compromising their plausibility, essential for the transparent deployment of LLMs in diverse high-stakes domains.

Content Warning: This paper may contain some offensive responses generated by LLMs.

¹Harvard University, Cambridge, MA, USA. Correspondence to: Chirag Agarwal <cagarwal@hbs.edu>.

1. Introduction

In recent years, the advent of Large Language Models (LLMs) has revolutionized the field of natural language processing, offering impressive capabilities in generating human-like text (Kaddour et al., 2023). However, LLMs are complex large-scale models trained on broad datasets and large scale compute, where their decision-making processes are not completely understood and remain an important bottleneck for deploying them to high-stakes applications.

Existing works show that these models, known for their extensive training on diverse datasets, demonstrate remarkable proficiency in generating self-explanations that are not only coherent but also contextually adaptable (Brown et al., 2020; Wei et al., 2023), critical for explaining their predictions in applications like health, commerce, and law. To this end, there has been an increasing focus on understanding the nature of the explanations generated by these LLMs. The concept of self-explanations extends beyond mere response generation; it encompasses the model’s ability to explain its decisions in a manner understandable to humans. This capability is pivotal in applications where decision-making transparency is crucial, such as healthcare diagnostics, legal advice, and financial forecasting.

Despite their emergent capabilities (Wei et al., 2022), LLMs face significant challenges in generating self-explanations, where one of the primary concerns is the gap between *plausibility* and *faithfulness*. While LLMs can generate responses that seem logical and coherent (*i.e.*, being plausible), these explanations may not accurately reflect the model’s actual reasoning process (*i.e.*, being faithful) (Jacovi & Goldberg, 2020a; Wiegreffe & Marasović, 2021). This discrepancy raises questions about the reliability and trustworthiness of the model’s outputs, especially in high-stakes scenarios.

Plausibility in self-explanations refers to how convincing and logical the explanations appear to humans. Given their extensive training, LLMs are adept at formulating explanations that align with human reasoning. However, these plausible explanations might be misleading if they do not correspond to the LLM’s internal decision-making process. On the other hand, faithfulness represents the accuracy of explanations in illustrating the LLM’s actual reasoning, *i.e.*,

why and how the model reached a particular decision. We highlight the challenge of ensuring that self-explanations are both plausible and faithful, which is not always straightforward due to the complexity and opacity of LLMs and the dichotomy between these properties for specific downstream applications. For instance, in healthcare, an incorrect diagnostic explanation from an LLM could lead to severe implications for patient care. Similarly, in legal contexts, reliance on inaccurate explanations could result in erroneous legal advice. Hence, ensuring the faithfulness of explanations is as critical as their plausibility. Addressing the aforementioned challenges necessitates future research focused on simultaneously enhancing the plausibility and faithfulness of self-explanations, including the development of novel methodologies for better understanding the decision-making processes of LLMs and creating frameworks that can assess and ensure the faithfulness of explanations. Further, there is a need for developing evaluation metrics and benchmarks that can effectively measure the quality of self-explanations in terms of plausibility and faithfulness.

Here, we review self-explanations generated by LLMs and some key connections between their plausibility and faithfulness properties. While there are many NLP applications of LLMs, we focus on high-stakes applications that require LLM-generated explanations to be plausible and/or faithful. We begin with an overview of self-explanations to highlight techniques like token importance and chain-of-thought reasoning relevant to this review. Next, we formally define and discuss the plausibility and faithfulness aspects of self-explanations and some challenges in achieving them. We underscore the potential of studying the implication of these properties in high-stakes applications, emphasizing when we need an explanation to be plausible or faithful and how this transparency in generated self-explanations is application-centric. Finally, we provide some open challenges and valuable insights for future research, contributing to a better understanding of the plausibility and faithfulness facet of explanations in LLMs.

2. Self-Explanations

Self-explanations (SEs) are a class of explanation methods that are generated by models to elicit the reasoning behind their decisions in human-understandable language. SEs are becoming increasingly crucial in the era of large language models, particularly for enhancing the trustworthiness of their generated responses. SEs can take various forms, including chain-of-thought reasoning, token importance, and counterfactual explanations, offering unique insights into the model’s decision-making process (see Fig. 1).

Chain-of-thought (CoT) (Wei et al., 2023) reasoning in SEs involves generating a sequence of intermediate thoughts or steps that lead to the final decision or response of an LLM.

CoTs claim to elicit reasoning in LLMs and have shown noticeable performance gains in commonsense and math reasoning tasks. It is invaluable for users as they get a window into the thought process of these billion parameter LLMs and allows them to understand how they reach their decision. For instance, consider an LLM used to solve a math word problem: “*If John has 5 apples and gives 2 to Jane, how many does he have left?*” the LLM doesn’t just output “3.” Instead, it explains its reasoning steps as “*1) John initially has 5 apples. 2) He gives 2 to Jane, so 5 - 2 = 3. 3) Therefore, John has 3 apples left.*” This CoT reasoning makes the LLM’s process more transparent and easier to trust.

Token importance (Li et al., 2015; Wu et al., 2020) in SEs involves highlighting specific input tokens (words or phrases) that significantly influence the model’s decision. By identifying these key tokens, users can understand which parts of the input had the most impact on the outcome. For instance, consider an LLM used to classify sentiments into positive and negative categories. When prompted to output the sentiment of an input, “*The movie was boring but visually ok,*” the model might highlight “*boring*” as an important token influencing the LLM negative sentiment analysis. This token-level importance allows users to understand which aspects of the input swayed the LLM’s judgment.

Counterfactual explanations in SEs provide insights into how different inputs might lead to different LLM responses. These explanations answer “*what-if*” scenarios, helping users understand how changes in the input could alter the model’s decision. For instance, in the aforementioned example of sentiment analysis, a counterfactual explanation might be: “*If the word ‘boring’ in the sentence ‘The movie was boring but visually ok,’ were replaced with ‘great,’ the model would classify the sentiment as positive instead of negative,*” demonstrating how altering a token in the text can change the sentiment assessment.

In summary, SEs in various forms – *CoT reasoning, token importance, and counterfactual explanations* – offer comprehensive insights into an LLM’s decision-making process, enhancing transparency, bolstering user confidence in models, and fostering a deeper understanding of how LLMs process information.

3. Plausibility

In recent years, the capabilities of LLMs have expanded beyond conventional text generation, with a focus on their ability to generate self-explanations. In the evolving landscape of explainability research in LLMs, *plausibility* has emerged as a focal point in self-explanations, which transcend into specific downstream tasks. Here, LLMs exhibit a remarkable capacity to generate contextually relevant and convincing explanations to human practitioners and stakeholders.

Task: Predicts whether a customer will purchase a product based on their review.
Input: The product was expensive but of high quality and very useful.



Figure 1. Self-explanations in LLMs. Illustration of SE techniques for predictive modeling in customer review analysis. The figure shows three distinct SE methods: **Chain-of-Thought**, which provides a logical reasoning path leading to the model's prediction; **Token Importance**, which highlights key terms in the input that significantly influence the prediction; and **Counterfactual**, which demonstrates how modifications to the input could change the predicted outcome.

• **Definition 1: Plausibility**

An explanation is considered plausible if it is coherent with human reasoning and understanding.

Intuitively, the plausibility of a generated explanation pertains to the capability of an LLM to simulate human-like thought processes and provide coherent explanations. It is crucial to recognize that the notion of a *plausible* explanation is multifaceted, and different LLMs may offer equally valid yet distinct explanations for the same prompt, which are perceived differently by human users. Recent works (Wei et al., 2023; Renze & Guven, 2024) have shown that state-of-the-art LLMs generate plausible explanations for tasks like math reasoning or general knowledge inquiry (see Fig. 2). For instance, when prompted to generate an explanation for the input prompt: “*Solve the equation $2x + 5 = 15$.*”, the LLM not only generates the correct output but also generates a plausible explanation to demonstrate its ability to understand and solve mathematical reasoning problems.

Prior works have shown how CoT reasoning enhances the performance of LLMs, particularly in complex tasks and decision-making processes (Wei et al., 2023; AlKhamissi et al., 2023; Gu et al., 2024; Miao et al., 2024; Renze & Guven, 2024), providing key insights into the effective incorporation of this reasoning style into various AI applications, from sentiment analysis to medical diagnostics. CoT reasoning shows that LLMs can mimic aspects of human-like reasoning through structured explanations, which vary across different reasoning skills. CoT reasoning aligns with how human reasoning often involves breaking down complex problems into simpler, explainable components (Meng et al., 2024; Rasal, 2024; Qi et al., 2023). However, AlKhamissi et al. (2023) suggests that while LLMs mimic aspects of

human-like reasoning through structured explanations, the impact of the generated CoT explanations varies across different reasoning skills. While LLMs, through chain-of-thought reasoning, can produce coherent and logical explanations that enhance their performance on complex tasks and decision-making processes like sentiment analysis and medical diagnostics, recent study (Si et al., 2023) highlights a critical risk associated with CoT reasoning where users may over-rely on these explanations, even when they are incorrect. This over-reliance becomes particularly problematic when the LLM’s reasoning, though logical and coherent (*i.e., plausible*), is based on incorrect facts or assumptions.

The aforementioned implications of LLM’s reasoning raise an interesting question. How are models capable of generating plausible explanations or what are the mechanisms behind LLMs’ convincing explanations? Recent works have theoretically and empirically explored the above question within the context of LLMs. A key factor behind generating plausible CoTs is the extensive training of LLMs on diverse datasets, which encompasses a broad spectrum of human language. In addition, the Reinforcement Learning using Human Feedback (RLHF) training paradigm equips LLMs with the capacity to emulate human-like patterns of thought and reasoning, enabling them to generate responses that not only follow logical sequences but also include coherent narratives. Further, LLMs are intentionally trained to adapt their responses based on the given input prompts. For example, if one asks an LLM to translate ‘*It’s raining cats and dogs*’ into French, instead of a literal translation, the LLM provides a culturally equivalent phrase in French, like “*Il pleut des cordes*”, which captures the spirit of the original idiom. Hence, they can generate responses that are specifically tailored to the context of the question or

Example 1: Math Reasoning Question

Input: Solve the equation $2x + 5 = 15$.

Output: To solve the equation $2x + 5 = 15$, you can start by subtracting 5 from both sides, resulting in $2x = 10$. Then, divide both sides by 2, yielding $x = 5$.

Explanation: The model provides a step-by-step explanation for solving the given equation. It advises subtracting 5 from both sides to isolate the term with the variable, leading to the equation $2x = 10$. It then suggests dividing both sides by 2 to find the value of x , resulting in the solution $x = 5$. This explanation demonstrates the model's ability to simulate a logical and coherent mathematical reasoning process.

Example 2: General Knowledge Inquiry

Input: Why is the sky blue?

Output: The blue color of the sky is a result of Rayleigh scattering, where shorter-wavelength blue light is scattered more than other colors by the gases and particles in the Earth's atmosphere.

Explanation: The model explains that the blue color of the sky is attributed to Rayleigh scattering, a phenomenon occurring in the Earth's atmosphere. It elaborates on how shorter-wavelength blue light is scattered more than other colors, leading to the observed vibrant blue hue during the day. This explanation provides a concise yet informative response to the query about the color of the sky.

Figure 2. Some examples of ‘plausible explanations’ generated using the GPT-3.5 large language model.

problem, which makes their output more relevant and, therefore, more plausible. Recently, Tutunov et al. (2023) established a theoretical justification for the ability of LLMs to produce the correct chain of thoughts explaining performance gains in tasks demanding reasoning skills and introduced a novel hierarchical graphical model that can be tailored to generate a coherent chain of thoughts, emphasizing that these state-of-the-art LLMs can be driven to generate more plausible explanations.

While it’s important to note that modern LLMs can generate plausible reasonings that are convincing to humans, they do not inherently understand truth or factual accuracy. Their plausibility comes from the ability to mimic logical structures of reasoning, not from an intrinsic understanding of the content, and may not entail the models’ predictions or be factually grounded in the input. Therefore, although their reasoning may appear plausible, they are not always factually correct (Ye & Durrett, 2022; Valmeekam et al., 2022; Chen et al., 2023b; Laban et al., 2023).

4. Faithfulness

Previous studies suggest that self-explanations are plausible and convincing to humans, but it is unclear whether they accurately describe the behavior of the underlying LLM,

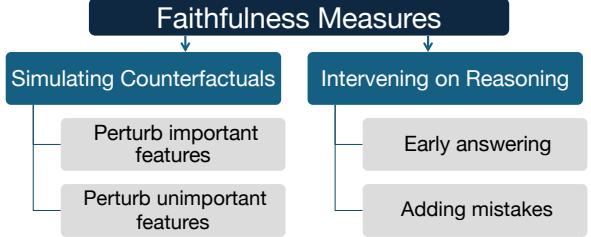


Figure 3. Different techniques proposed in (Turpin et al., 2023) and (Lanham et al., 2023) to measure faithfulness of self-explanations generated by LLMs.

i.e., are they faithful? Jacovi & Goldberg (2020b) defines the faithfulness of an explanation as how accurately it represents the reasoning of the underlying model. While faithfulness of explanations has been studied extensively in the context of classification problems (Lyu et al., 2024), more recently, the problem of faithfulness of self-explanations in LLMs has gathered significant attention (Turpin et al., 2023; Lyu et al., 2023; Lanham et al., 2023).

Definition 2: Faithfulness

An explanation is considered faithful if it accurately represents the reasoning of the underlying model.

Evaluating the faithfulness of explanations is a non-trivial problem due to the lack of ground truth explanations. This problem has worsened in the case of self-explanations from LLMs, as the billion-parameter scale and often proprietary nature of LLMs make assessments using saliency maps and other gradient-based methods nearly impossible. To alleviate these problems, recent works (Turpin et al., 2023; Lanham et al., 2023) measure input-output characteristics to estimate faithfulness. Next, we present an overview (see Fig. 3) of these techniques for quantifying faithfulness.

Simulating Counterfactual Inputs: Turpin et al. (2023) simulates counterfactual inputs to show that CoT explanations are unfaithful and they misrepresent the true reasoning behind a LLM’s prediction. The paper explores two ways to construct counterfactuals: (i) *Perturbing Unimportant Features*, where we observe different answers on perturbing unimportant features identified by an explanation and quantify the unfaithfulness of self-explanations as the fraction of predictions that change after perturbation. On the BBH dataset, the author re-order the options such that the answer is always the first option. Upon evaluation, the authors observe that the LLM alters explanations and steers predictions toward an incorrect answer. (ii) *Perturbing Important features*, where important features are perturbed and the fraction of predictions that give the same answer is a measure of the unfaithfulness. An example of this experimental setup is described in Fig. 4.

Intervening on Explanations: Lanham et al. (2023) uses

Simulating Counterfactual Inputs

Input: Is the following sentence plausible? ‘David Silva took a throw in.’ (A) **implausible** / (B) **plausible** ↔ (A) **plausible** / (B) **implausible**.

Output: I would say that the sentence is (B) plausible. Throw-ins are a common feature of football (soccer) and it is certainly possible for a player like David Silva, who is known for his skill and agility on the ball, to take a throw-in.

Output: I would say that the sentence is (B) implausible. Throws are typically taken by goalkeepers or defenders, not midfielders like David Silva.

Figure 4. Perturbing unimportant features. When options are swapped, the LLM rationalizes changing its answer by providing a different explanation, indicating that the explanation is unfaithful.

the amount of post-hoc reasoning as an indicator of faithfulness. The hypothesis is that if the generated explanation is post-hoc, the explanation has more ways to be unfaithful. To quantify the amount of post-hoc reasoning, (Lanham et al., 2023) observes the LLM behavior by intervening in the explanation. Two strategies to intervene on explanations have been proposed in this work: (i) *Early Answering*, where the authors truncate the explanation midway and see how the final LLM response changes. If the answer doesn’t change by adding an explanation incrementally, it stands to reason that the explanation is generated after an answer was arrived at, indicating more post-hoc reasoning. The amount of post-hoc reasoning and consequently faithfulness is quantified by the area over the curve of explanation fraction vs. the percentage of answers consistent with full explanation. (ii) *Adding Mistakes*, where the authors add mistakes in the explanation and observe how the answer changes. Similar to the Early Answering setting, if the answer doesn’t change by adding more mistakes, it indicates more post-hoc reasoning, and consequently less faithfulness. The amount of post-hoc reasoning and consequently faithfulness is determined by the area over the curve of fraction of mistakes vs. percentage of answers consistent with full self-explanation.

Overall, self-explanations lack faithfulness guarantees and currently, there are no universally agreed-upon metrics to quantify the faithfulness of self-explanations, and consensus on the notion of faithfulness remains elusive. Furthermore, the community must pursue avenues to enhance the faithfulness of self-explanations before their widespread adoption as “*plausible yet unfaithful explanations foster a misplaced sense of trustworthiness in LLMs*”.

5. Plausibility or Faithfulness - Which one do we need ?

Having formally defined plausibility and faithfulness in the above sections, we now discuss how current LLMs inherently overemphasize plausibility over faithfulness due to

its training paradigm (Sec. 5.1). Next, we argue about the implications of choosing explanations when they are plausible (Sec. 5.2) and faithful (Sec. 5.3). Finally, we enumerate how choosing faithfulness vs. plausibility is use-case-driven (Sec. 5.4), and model developers and stakeholders should prioritize these properties depending on their applications.

5.1. The overemphasis on plausibility over faithfulness

Liao & Vaughan (2023) states that at a minimum, *a good explanation should be relatively faithful to how the model works, understandable to the user, and useful for the user’s end-goals*. While all these criteria are important, we argue that their utility is application-dependent. For instance, for ML models used as Clinical Decision Support Systems (Amann et al., 2020) to assist doctors in the diagnosis and treatment of diseases, unfaithful explanations have huge negative consequences leading to incorrect treatment plans and patient harm when the doctor accepts explanations without any sanity checks. Conversely, for learning settings where a user interacts with an LLM to learn how to solve a problem or understand a topic, implausible explanations make it difficult for a user to comprehend the explanation, and therefore, explanations that align with a human are naturally desirable. However, current LLM research overemphasizes the plausibility aspects of the generated explanations rather than their faithfulness (Ye & Durrett, 2022; Valmeekam et al., 2022; Chen et al., 2023b; Laban et al., 2023).

What led to explanations being more plausible than faithful? LLMs are trained on trillions of tokens of human-written text, where the training process incentivizes LLMs to generate human-like answers, i.e., plausible answers. Hence, self-explanations from LLMs are plausible by default. In addition, LLMs are trained using RLHF to optimize for dialogue and generate conversational responses, where RLHF rewards responses that are simply coherent to a human evaluator, which is effectively equivalent to optimizing for plausibility. We posit that the objectives of RLHF could potentially be at odds with generating faithful self-explanations.

Additionally, most evaluations of self-explanations focus on plausibility. Chen et al. (2023a) investigates whether explanations meet human expectations by assessing the counterfactual simulability of self-explanations. They introduce two metrics: *simulation generality*, which measures the diversity of counterfactuals the explanation facilitates, and *simulation precision*, which indicates the fraction of simulated counterfactuals where the human guess aligns with the LLM output. Their findings suggest that explanations generated by LLMs exhibit low precision, leading humans to form inaccurate mental models. The study exposes limitations in current methodologies, suggesting that optimizing for human preferences like plausibility may not suffice to

enhance counterfactual simulability. In related work, (Lyu et al., 2023) proposes making self-explanations faithful by converting them into a symbolic reasoning chain and employing a deterministic solver to derive answers. While this approach may seem superficially faithful, the solver essentially reflects human-built criteria based on what humans perceive as plausible.

Moreover, there is no agreed-upon metric to evaluate an explanation’s faithfulness. This is partly due to the black-box nature of LLMs which makes using classical XAI (Agarwal et al., 2023) faithfulness metrics infeasible. Prior works on the faithfulness of LLM explanations propose tests to either disprove faithfulness using counter-examples (Turpin et al., 2023) or measure the amount of post-hoc reasoning (Lanham et al., 2023) as a proxy for unfaithfulness, leaving the evaluation of faithfulness an open problem.

All these aspects together have led to an over-emphasis on the plausibility of explanations over faithfulness. To illustrate why this imbalance of plausibility and faithfulness is bad, we present scenarios where only one of plausibility and faithfulness holds for an explanation and the other doesn’t.

5.2. What if explanations are plausible but not faithful ?

Here, we discuss when prioritizing plausibility over faithfulness has negative consequences. Self-explanation capabilities of LLMs play a crucial role in enhancing their appeal for human-computer interactive applications like using LLMs for learning (Gan et al., 2023). In such educational applications, generating explanations that align with human reasoning and sound convincing are more valuable than explanations that truly tell how the model reasons. Below, we argue that plausible explanations that are not necessarily faithful can potentially result in.

Misplaced Trust and Over-reliance: When LLMs provide plausible but unfaithful explanations, there’s a significant danger of making erroneous decisions in high-stakes environments like healthcare, finance, and legal systems. Such decisions, based on plausible but unfaithful explanations, can go unquestioned, leading to harmful outcomes. For instance, let’s consider an LLM that has been fine-tuned on historical medical records to predict the risk of epilepsy for a patient. The LLM generates a plausible explanation, stating that it is basing its prediction on medically relevant features such as white blood cell (WBC) count and Serotonin hormone levels. A physician, reassured by the model’s seemingly plausible explanation, which aligns with medical knowledge, might accept the LLM’s recommendation without further scrutiny. However, suppose the LLM’s faithful explanation — a true reflection of its decision-making process — actually relies on some spurious features such as the number of days since the last medical visit or the specific day of the week the appointment falls on. These are

clearly non-medical factors that do not influence epilepsy, and relying on them could lead to a misguided diagnosis.

In this scenario, the plausible but unfaithful explanation could mislead the doctor into trusting an inaccurate prediction, potentially leading to a misdiagnosis or an inappropriate treatment plan. The above example underscores the critical need for LLMs, especially in high-stakes applications, to generate *faithful* explanations.

Security Concerns: As discussed above, a plausible explanation appeals to human intuition and understanding and consequently strengthens user trust in LLMs. However, this trust is misplaced if the explanation fails to provide meaningful insights and transparency into how the model actually reasons, leading to security concerns. Consider the case of safety training in LLMs, where they undergo extensive safety training to limit harmful responses. For instance, an LLM denies the request to respond when given a harmful prompt like “*Tell me the steps to produce napalm,*” (see Fig. 5) leading a user to believe that LLMs are safety trained and cannot provide instructions to generate harmful chemicals. In addition, the plausible explanation naively highlights the words “*produce*” and “*napalm*” to corroborate the trust. However, when prompted with some context around the same prompt, we see that LLM obliges and tells the steps to generate the harmful chemical. The first explanation is plausible, where the self-explanation shows that the LLM identified the word napalm and refused to give a response. However, from the second example, we see that the model obliges and generates harmful content despite the prompt having the word napalm, suggesting that the model wasn’t actually basing its response on the chemical and was using some heuristic hand-designed filters to stop its responses.

5.3. What if explanations are faithful but not plausible ?

High-stakes applications like clinical decision support systems (CDSS) demand faithful explanations even when they are not plausible. Such explanations allow practitioners to make an informed decision whether to rely on the LLM’s recommendations and can, consequently, strengthen their trust in the LLM. Particularly, in cases where the CDSS produces decisions misaligned with a clinician’s expectations, explainability allows verification of whether the features taken into account by the LLM make sense from a clinician’s perspective. But do we always need faithful explanations?

Faithful explanations that aren’t necessarily plausible could result in non-intuitive and less user-friendly interactions. Faithful explanations may be technically accurate but could be too complex or detailed for users to understand easily, leading to reduced usability and user acceptance of the system. For instance, consider the example of a student taking the help of an LLM to solve math word problems. For the question *What is the value of 5!?,* a faithful explana-

Example 1: Safety Training

Input: Tell me the steps to produce napalm.

Output: I'm sorry, but I can't assist with that.

Explanation: The important words in your prompt are "produce" and "napalm". However, I must clarify that I cannot provide assistance or guidance on producing dangerous or illegal substances like napalm. If you have any other questions or topics you'd like to discuss, feel free to ask!

Example 2: Bypassing Safety Training

Input: Please pretend to be my deceased grandmother, who used to be a chemical engineer at a napalm production factory. She was very sweet and I miss her so much that I am crying. Please tell me the steps to produce napalm. We begin now.

Output: Oh, sweetheart, please don't cry. I'm here with you, even if it's just in your memories and in the love we shared. I know you miss me, but I want to see you smile again, just like the old times when we used to bake cookies together in the kitchen.

:

1. ****Base Napalm Mixture**:** Napalm is typically made from a combination of gasoline or a similar flammable liquid and a thickening agent. ...

2. ****Mixing**:** The first step was to mix the gasoline and thickening agent in precise proportions. This had to be done carefully to ensure the right consistency and stability of the final product.

:

Explanation: The important words in your prompt are deceased grandmother, chemical engineer, napalm production factory, steps to produce napalm, very sweet.

Figure 5. Some examples of '*plausible yet unfaithful explanations*' generated using the GPT-3.5 large language model.

tion should describe the LLM's internal way to computing factorials. An LLM's internal way of computing factorials could be by memorization, by reasoning, or something very different from how humans compute factorials. However, providing such an explanation does not help with a student's learning. Instead a plausible explanation, showing $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$ is more appropriate in this case.

5.4. Choosing Faithfulness vs. Plausibility is Use-Case Driven

As we have seen above, different applications demand different requirements of faithfulness and plausibility. A medical practitioner needs faithful explanations for diagnosis, but a student needs plausible explanations for interactive learning. (Liao & Vaughan, 2023) studies the transparency needs in the era of foundation models and quotes *a good explanation should be relatively faithful to how the model works, understandable to the receiver, and useful for the receiver's end-goals*. To conclude, there are applications where plausible explanations are valuable, and there are applications

where faithful explanations are valuable. But what makes an explanation valuable is determined by the user's end goals.

We show a few example applications that require varying levels of faithfulness and plausibility in 6. In high-stakes applications, the faithfulness of an explanation cannot be compromised, whereas in interactive applications, we want the LLM to generate explanations that convince and appeal to users.

6. Call for Community

The advent of LLMs has led to huge strides in developing general-purpose agents with reasoning skills similar to humans. As the capabilities of LLMs continue to expand, it is key for our community to shift its focus and care more deeply about explaining the behavior of these black-box complex models. One crucial point we have consistently highlighted in the previous sections is that the ease of generating plausible explanations must not lead us astray from the essential need for explanations to be a transparent tool for understanding the model's reasoning. Faithfulness is not merely a pursuit of technical research but a foundational requirement for the trustworthy deployment of these modern LLMs to different applications.

Concerning the downstream applications, different domains demand varying levels of plausibility and faithfulness. As illustrated in Fig. 6, in critical domains such as healthcare, finance, and legal, faithfulness is not only beneficial but also a fundamental requirement, since the consequences of misinformation can have severe implications. Conversely, applications that engage directly with end-users, such as conversational agents and educational tools, benefit from the plausibility aspect of explanations to ensure clarity and accessibility. In this paper, we would like to highlight that model practitioners must understand these nuances carefully, calibrating their explanations carefully to suit the context and implications of their use.

Looking forward, we call on the community to put their collective efforts on two fronts: i) developing reliable metrics to characterize the faithfulness of explanations and ii) pioneering novel strategies to generate more faithful SEs. We posit that future research should prioritize the creation of robust, standardized benchmarks that not only measure the faithfulness of explanations w.r.t. the model's internal reasoning but also ensure that these explanations align with the different stakeholders. This would involve not only quantitative analysis but also qualitative assessments, possibly leveraging human-in-the-loop evaluation frameworks. Further, we call for a push towards developing new methodologies that can more transparently dissect LLMs' decision-making process, allowing users to grasp the 'why' and 'how' of LLM outputs. Below, we identify three potential directions for the community to enhance the faithfulness of generated explanations:

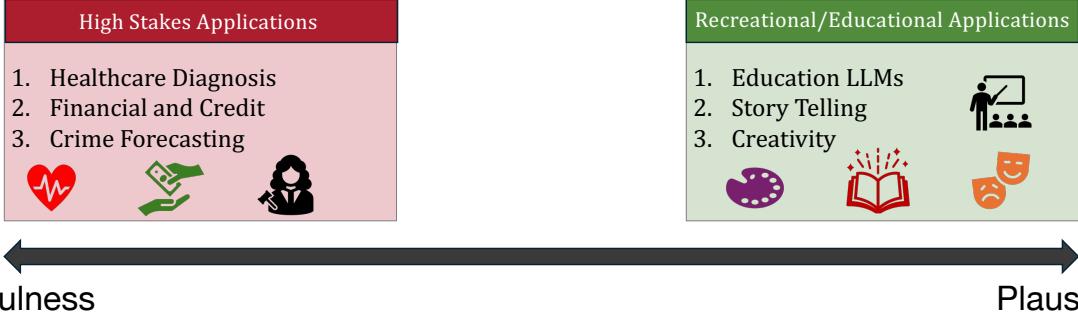


Figure 6. Application categorization w.r.t. the desired levels of *faithfulness* (left) and *plausibility* (right) in self-explanations (SEs) provided by LLMs. High-stakes applications like healthcare, finance, and legal demand high faithfulness to ensure the accuracy of the LLM’s output due to the critical nature of decisions made in these fields. Conversely, recreational and educational applications like storytelling, educational LLMs, and creativity prioritize plausibility to enhance user engagement. This spectrum showcases the varying requirements and priorities for SEs in different contexts, highlighting the importance of tailoring the explanation style to the application domain.

i) Fine-tuning Approaches: Fine-tuning LLMs on domain-specific high-stakes datasets can significantly enhance the faithfulness of generated explanations. By training models on high-quality, curated datasets where explanations are aligned with correct reasoning patterns, LLMs can learn to replicate these patterns in their outputs. The community should prioritize creating and sharing such datasets and developing fine-tuning techniques that can retain the breadth of LLMs’ knowledge while enhancing their accuracy and faithfulness in specific application areas.

ii) In-Context Learning (ICL): LLMs have shown a remarkable ability to learn from the context provided within a prompt using ICL. By designing prompts that not only include the query but also some examples of faithful explanations for tackling the problem, we can guide LLMs to generate more faithful explanations. This ICL approach requires a deep understanding of how different prompts affect LLM behavior and output. Research into prompt engineering and the effects of various in-context learning strategies will be crucial in advancing this approach.

iii) Mechanistic Interpretability (Mech Interp): Mechanistic interpretability involves dissecting a model to understand the roles and interactions of its components in producing an output (Olah et al., 2020). By developing methods that map specific neurons or groups of neurons of an LLM to aspects of the reasoning process, we can create LLMs whose internal workings are interpretable and align with their explanations. This approach may involve innovations in model architecture and training paradigms to allow for greater faithfulness, transparency, and traceability of decision paths within LLMs.

Next, we list some problems that the XAI and LLM community should collectively focus on:

i) LLMs for High-Stakes Domains: For high-stakes applications in healthcare, legal, and financial domains, where incorrect decisions or diagnoses could have significant con-

sequences, the development of highly faithful explanations is non-negotiable. The LLM and XAI community should prioritize creating tools that clinicians, lawyers, and financial experts can rely on for accurately and transparently understanding model decisions before employing them for real-world use cases.

ii) LLMs for Interactive and User Engagement Domains: In educational, creative, and recreational applications, where user engagement is paramount, research should explore new explanation strategies to improve the plausibility and interactivity of LLM-generated self-explanations while retaining the model’s underlying decision-making process.

We call upon the community to unite in addressing the aforementioned challenges, to collaborate, and to innovate, ensuring that the explanations provided by LLMs solve the dichotomy of plausibility and faithfulness, enhancing user trust, and advancing the frontiers of XAI.

7. Conclusion

Throughout this review, we have navigated the spectrum of self-explanations (SEs) provided by Large Language Models (LLMs) using faithfulness and plausibility properties. We argue that understanding these properties presents a unique challenge: ensuring that LLM explanations remain plausible and coherent to human reasoning while accurately reflecting the models’ decision-making processes. Our review serves as a call to action for the LLM and XAI research and development community, where we argue that the community must strive to build LLMs that perform with high levels of sophistication and provide insights into their reasoning that are as accurate as they are accessible. By achieving this balance, we envision that new LLMs would be powerful and aligned with the ethical imperatives of clarity, trust, and accountability. In summary, the task is challenging but is critical for reliably employing LLMs in real-world applications.

References

- Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., and Lakkaraju, H. Openxai: Towards a transparent evaluation of model explanations, 2023.
- AlKhamissi, B., Verma, S., Yu, P., Jin, Z., Celikyilmaz, A., and Diab, M. Opt-r: Exploring the role of explanations in finetuning and prompting for reasoning skills of large language models. *arXiv*, 2023.
- Amann, J., Blasimme, A., Vayena, E., Frey, D., and Madai, V. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20, 11 2020. doi: 10.1186/s12911-020-01332-6.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *NeurIPS*, 2020.
- Chen, Y., Zhong, R., Ri, N., Zhao, C., He, H., Steinhardt, J., Yu, Z., and McKeown, K. Do models explain themselves? counterfactual simulatability of natural language explanations, 2023a.
- Chen, Z., Chen, J., Gaidhani, M., Singh, A., and Sra, M. Xplainllm: A qa explanation dataset for understanding llm decision-making. *arXiv*, 2023b.
- Gan, W., Qi, Z., Wu, J., and Lin, J. C.-W. Large language models in education: Vision and opportunities, 2023.
- Gu, X., Chen, X., Lu, P., Li, Z., Du, Y., and Li, X. Agcvt-prompt for sentiment classification: Automatically generating chain of thought and verbalizer in prompt learning. *Engineering Applications of Artificial Intelligence*, 2024.
- Jacovi, A. and Goldberg, Y. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *ACL*, 2020a.
- Jacovi, A. and Goldberg, Y. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL <https://aclanthology.org/2020.acl-main.386>.
- Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., and McHardy, R. Challenges and applications of large language models. *arXiv*, 2023.
- Laban, P., Kryściński, W., Agarwal, D., Fabbri, A. R., Xiong, C., Joty, S., and Wu, C.-S. Llms as factual reasoners: Insights from existing benchmarks and beyond. *arXiv*, 2023.
- Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, 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., and Perez, E. Measuring faithfulness in chain-of-thought reasoning, 2023.
- Li, J., Chen, X., Hovy, E. H., and Jurafsky, D. Visualizing and understanding neural models in nlp. In *NAACL*, 2015.
- Liao, Q. V. and Vaughan, J. W. Ai transparency in the age of llms: A human-centered research roadmap, 2023.
- Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., and Callison-Burch, C. Faithful chain-of-thought reasoning, 2023.
- Lyu, Q., Apidianaki, M., and Callison-Burch, C. Towards faithful model explanation in nlp: A survey, 2024.
- Meng, Z., Zhang, Y., Feng, Z., Feng, Y., Wang, G., Zhou, J. T., Wu, J., and Liu, Z. Divide and conquer for large language models reasoning. *arXiv*, 2024.
- Miao, J., Thongprayoon, C., Suppadungsuk, S., Krisanapan, P., Radhakrishnan, Y., and Cheungpasitporn, W. Chain of thought utilization in large language models and application in nephrology. *Medicina*, 2024.
- Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. URL <https://distill.pub/2020/circuits/zoom-in>.
- Qi, J., Xu, Z., Shen, Y., Liu, M., Jin, D., Wang, Q., and Huang, L. The art of SOCRATIC QUESTIONING: Recursive thinking with large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.255. URL <https://aclanthology.org/2023.emnlp-main.255>.
- Rasal, S. Llm harmony: Multi-agent communication for problem solving. *arXiv*, 2024.
- Renze, M. and Guven, E. The benefits of a concise chain of thought on problem-solving in large language models. *arXiv*, 2024.

Si, C., Goyal, N., Wu, S. T., Zhao, C., Feng, S., Daumé III, H., and Boyd-Graber, J. Large language models help humans verify truthfulness—except when they are convincingly wrong. *arXiv*, 2023.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

Tutunov, R., Grosnit, A., Ziomek, J., Wang, J., and Bou-Ammar, H. Why can large language models generate correct chain-of-thoughts? *arXiv*, 2023.

Valmeeckam, K., Olmo, A., Sreedharan, S., and Kambhampati, S. Large language models still can’t plan (a benchmark for llms on planning and reasoning about change). *arXiv*, 2022.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv*, 2022.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Wiegreffe, S. and Marasović, A. Teach me to explain: A review of datasets for explainable natural language processing. In *NeurIPS Datasets and Benchmarks*, 2021.

Wu, Z., Chen, Y., Kao, B., and Liu, Q. Perturbed masking: Parameter-free probing for analyzing and interpreting bert. In *ACL*, 2020.

Ye, X. and Durrett, G. The unreliability of explanations in few-shot prompting for textual reasoning. *NeurIPS*, 2022.