Check for updates

REVIEW

Comparison of explainability methods for hallucination analysis in LLMs

[version 1; peer review: 1 approved with reservations, 2 not approved]

Ioannis Papagiannopoulos, Hercules Koutalidis, Panagiota Rempi, Christos Ntanos, Dimitrios Askounis

School of Electrical and Computer Engineering, National Technical University of Athens, Zografou, Attica, 15773, Greece

v1 **First published:** 23 Jul 2025, **5**:191
https://doi.org/10.12688/openreseurope.20839.1
Latest published: 23 Jul 2025, **5**:191
https://doi.org/10.12688/openreseurope.20839.1

Abstract

The primary obstacle to the safe application of artificial intelligence (AI) in sensitive fields such as healthcare and law is the phenomenon of Large Language Models (LLM) hallucinations, which produce fluent yet factually incorrect or illogical outputs. This paper investigates hallucinations as reasoning failures, distinguishing between factual inaccuracies and flawed logical derivations. We examine the limitations of current interpretability tools including Retrieval-Augmented Generation (RAG), attention mechanisms, SHapley Additive exPlanations (SHAP), and Local Interpretable Model-agnostic Explanations (LIME), which often fail to identify the root causes of these errors. To address this gap, we analyze reasoning in LLMs through two complementary paradigms: deterministic rule-based and stochastic data-driven approaches. We argue that hybrid reasoning techniques can improve the accountability and transparency of model behavior. A real-world use case in decision support systems illustrates the risks associated with transparent reasoning in high-stakes applications. We also present a comparative framework to evaluate the ability of common explainability techniques to detect and account for hallucinated outputs. Finally, we suggest useful techniques to improve trustworthiness, such as training data attribution, confidence calibration, and prompt-based abstention mechanisms. Our results highlight how crucial reasoning-aware explainability is as a starting point for ensuring that LLMs behavior complies with safety, dependability, and responsible AI deployment standards.

Plain Language Summary

Sometimes, ChatGPT and other large language models (LLMs) produce text that appears accurate but is inaccurate or deceptive. These errors, known as hallucinations, can lead to issues in fields such

Open Peer Review

Approval Status ✗ ✗ ?

	1	2	3
version 1 23 Jul 2025	✗ view	✗ view	? view

1. **Florian Leiser**, Technical University Munich, Munich, Germany
2. **Albert Sánchez Graells**, Bristol University, Bristol, UK
3. **Burak Yaprak**, İstanbul Ticaret University,, İstanbul, Turkey

Any reports and responses or comments on the article can be found at the end of the article.

as law or medicine. This study examines the reasons behind these mistakes made by LLMs as well as ways to better comprehend and account for them. To make these models more transparent and reliable, we propose combining data-driven reasoning with logical rules.

Keywords
Large Language Models, Hallucinations, Explainability, Reasoning, Trustworthy AI, Hybrid Reasoning, Decision Support Systems



This article is included in the Horizon Europe gateway.

Corresponding author: Ioannis Papagiannopoulos (ipapagiannopoulos@epu.ntua.gr)

Author roles: Papagiannopoulos I: Conceptualization, Investigation, Methodology, Project Administration, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Koutalidis H**: Conceptualization, Writing – Review & Editing; **Rempi P**: Conceptualization, Writing – Review & Editing; **Ntanos C**: Project Administration, Supervision, Writing – Review & Editing; **Askounis D**: Funding Acquisition, Supervision

Competing interests: No competing interests were disclosed.

Grant information: This work has been funded by the European Union's Horizon Europe research and innovation program under the ENERSHARE project, grant agreement No. 101069831

Copyright: © 2025 Papagiannopoulos I *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Papagiannopoulos I, Koutalidis H, Rempi P *et al.* **Comparison of explainability methods for hallucination analysis in LLMs [version 1; peer review: 1 approved with reservations, 2 not approved]** Open Research Europe 2025, **5**:191 https://doi.org/10.12688/openreseurope.20839.1

First published: 23 Jul 2025, **5**:191 https://doi.org/10.12688/openreseurope.20839.1

Introduction

On a variety of natural language processing tasks, Large Language Models (LLMs) like Claude, PaLM, and GPT-4 have demonstrated exceptional performance. They have been incorporated into high-impact applications, such as decision support systems in crucial fields like public administration, healthcare, and law, due to their capacity to produce logical, context-aware text[1]. There is however still a persistent and unsolved problem in spite of these advancements. These models frequently generate hallucinations, which are grammatically and syntactically correct, but factually incorrect, deceptive, or completely made up outputs[2].

The dependability and credibility of LLMs are seriously compromised by hallucinations. This is particularly problematic in fields where the generated text has the potential to directly affect human judgment. A hallucinated output can spread false information or result in less-than-ideal, even dangerous, outcomes in situations that call for accuracy and accountability[3]. Despite the prevalence and consequences of hallucinations, little is known about their underlying causes, and there are no systematic approaches to recognizing or elucidating them[4].

A number of explainability strategies have been put forth to deal with deep learning systems' opacity[5]. Nevertheless, the majority of these techniques are not appropriate for generative models such as LLMs[6]. Methods like saliency maps, attention visualizations, SHapley Additive exPlanations (SHAP), and Local Interpretable Model-agnostic Explanations (LIME) concentrate on token attributions at the surface level and frequently lack causal insight[7,8]. These tools usually work after the fact and infrequently show how the dynamics of decoding or the training data lead to inaccurate results. When trying to interpret hallucinated responses, their narrow scope is especially noticeable[9].

We suggest that rather than being viewed as outward anomalies, hallucinations should be viewed as reasoning errors. From this vantage point, we investigate whether adding structured reasoning paradigms to explainability initiatives can provide more insight into the causes and timing of hallucinations[10,11]. We investigate the possibilities of merging two complementary methods. While stochastic reasoning captures data-driven associations learned during training, deterministic reasoning reflects logic-based inference using predefined rules and constraints. A more thorough interpretation of model behavior might be possible with a reasoning-aware framework that incorporates both viewpoints[12].

This paper has four main objectives. First, we summarize recent findings related to hallucinations in LLMs. Second, we evaluate the effectiveness of current interpretability techniques through the lens of reasoning. Third, we examine how deterministic and stochastic reasoning modes intersect with explainability, and whether a hybrid approach can improve model transparency[13]. Fourth, we use an actual use case involving decision support systems to illustrate the usefulness of these concepts. We wrap up by comparing the best explainability tools and evaluating how well they can explain hallucinations. By conducting this study, we hope to support continued efforts to develop more reliable and responsible language technologies[14].

Background and key concepts

Clear conceptual foundations are essential to understand hallucinations in large language models. This section defines the key concepts and terminology that shape our analysis.

We begin by clarifying what hallucination means in the context of LLMs and how different types appear in outputs. Then, we discuss explainability in AI with a focus on methods applicable to generative models. Finally, we present deterministic and stochastic reasoning as two contrasting but potentially complementary approaches that influence model interpretability[15].

In LLMs, a hallucination refers to text that is fluent and grammatically correct but factually incorrect, misleading, or entirely fabricated. Unlike simple classification mistakes, these outputs often appear plausible, which makes them harder to detect. We identify two main categories[16]. Factual hallucinations involve incorrect claims about the real world, such as saying that the Eiffel Tower is in Berlin. Reasoning-based hallucinations occur when the model combines valid information but draws flawed conclusions. These categories can overlap and are especially concerning in settings that require high levels of trust and accuracy[17].

Explainability in generative AI models

Explainability techniques aid in our comprehension of how a model generates a particular result. These techniques are typically separated into two categories. After the model has produced its output, post-hoc techniques are used. Saliency visualizations, attention maps, SHAP, and LIME are a few examples. These tools frequently indicate which aspects of the input had an impact on the response, but they don't go into detail about how the model came to that conclusion[18]. On the other hand, intrinsic explainability methods are built into the model itself. They aim to make the decision-making process more transparent during generation rather than after it. Another way to categorize explainability is by scope. Local methods explain individual predictions, while global methods address general model behavior. Most LLM-focused techniques are local and post-hoc, and often struggle to reveal the roots of hallucinated content[15,19].

Logical vs. probabilistic reasoning in LLM hallucinations

To understand the sources of hallucinations, we must consider the reasoning processes that LLMs implicitly carry out. Although these models are not explicitly designed to reason, they often exhibit emergent reasoning-like behaviors such as pattern completion, analogical inference, and simple deductive sequences, despite lacking formal mechanisms for structured logic or rule chaining[20]. These can be grouped into two categories. Deterministic reasoning involves rule-based inference and draws logical conclusions from given premises[21]. It is typical of symbolic AI and systems built on structured knowledge. Stochastic reasoning, on the other hand, reflects probabilistic associations learned from data. This is how LLMs generally operate, predicting

tokens based on learned distributions without enforcing logical consistency[22]. Hallucinations often arise from failures in this probabilistic reasoning, such as invalid generalizations or unsupported associations[23].

There is a close connection between reasoning paradigms, explainability tools, and types of hallucinations. Many hallucinations are the result of flawed reasoning, and if the reasoning is hidden, interpretability tools may falsely validate incorrect outputs. The next section expands on this view by offering a detailed classification of hallucinations and examining their technical and behavioral causes[16].

Types and causes of hallucinations in LLMs

Hallucinations in large language models are not a single error type but a range of failures linked to different underlying issues. Understanding these distinctions is essential for evaluating their impact and for designing methods to prevent or explain them. Broadly, hallucinations fall into two categories: factual and reasoning-based[16].

Factual hallucinations

Factual hallucinations occur when the model generates information that is objectively incorrect. These are often easy to identify after the fact, such as claiming that the Eiffel Tower is in Berlin or referring to fictional events or people[24]. Such errors may result from gaps or biases in the training data, or from faulty associations learned during training. Since LLMs do not verify their outputs against external sources, they may confidently produce content that sounds plausible but is untrue[25].

Reasoning-based hallucinations

Logic or inference errors are a feature of reasoning-based hallucinations. The conclusion is flawed even though the facts are correct. For instance, a model might claim that whales are mammals and that all mammals give birth to live young, but it might draw the erroneous conclusion that whales do not lay eggs like birds. Although the language appears coherent, the reasoning behind it is invalid. These errors are harder to detect because they follow the surface rules of language[26]. It has been statistically proven that if a language model is well-calibrated, it must hallucinate certain types of facts[27].

Contributing factors and causes

Training data is the source of many hallucinations. LLMs use enormous datasets having accurate information for learning. Inaccurate details are in these datasets also. Without explicit filtering, the model may treat repeated patterns as reliable. Even without correction, the repeated patterns may still be treated as reliable by the model[28]. shows that this can lead to some errors with high confidence.

Prompt design is another key factor. As revealed in studies of prompt-induced bias as well as ambiguous prompts[28,29], the model is able to misinterpret user intent through vague or poorly structured prompts. Strategies in regard to sampling also do indeed matter. Techniques such as high temperature or aggressive beam search can prioritize fluency over factual accuracy, and also this increases the risk for hallucinated output[30].

Lack of grounding and external verification

A lack of the grounding in external, trusted sources heightens all hallucination risks. LLMs rely just on internal representations learned in training. They do not happen to have retrieval-augmented capabilities at all. Often, plausible-sounding content lacking supporting evidence results[31].

Hallucinations can be understood as the result of both model-level uncertainty and noise in the training data. Their form, whether factual or reasoning based, depends on the type of failure. However, their causes are often linked to training dynamics, prompt design, and output generation methods. Recognizing these patterns is the first step toward building more effective explanation techniques. We explore this direction in the following section[32].

Explainability techniques and their limitations

Overview of explainability methods

Several explainability methods have been introduced to interpret machine learning model outputs, with many adapted for LLMs. SHAP and LIME are widely used, offering local, post-hoc explanations by assigning importance to input features. Attention maps and saliency visualizations reveal which input parts the model focuses on during processing. Prompt-based justifications, by contrast, encourage the model to generate selfexplanations along with its answers[33,34].

Limitations in handling hallucinations

When applied to hallucinated outputs, these methods have significant limitations even though they provide some insight as they often fail to establish causal attribution or represent the influence of training data[35]. The majority of them only offer superficial attribution, failing to account for the underlying logic or the impact of training data. Furthermore, they don't provide a causal framework or an explanation for the model's particular error. Crucially, they struggle with the probabilistic nature of generation in LLMs, where latent patterns discovered from data also influence outputs in addition to input[36,37].

The role of Retrieval-Augmented Generation (RAG)

RAG systems operate by first retrieving passages or relevant documents and then generating responses from an external corpus conditioned on that context. This mechanism introduces explicit grounding in it. It does reduce some reliance upon the model's flawed internal representations[38]. LLMs and retrieval merge yielding grounded responses from reliable sources. RAG systems that happen to be well-designed return an answer plus identifiers such as URLs or else document IDs and excerpts coming from supporting documents[39]. This structure improves transparency, and users can verify the factual basis of the output.

RAG Limitations and open challenges

Even RAG systems remain vulnerable to hallucinations. These can occur when retrieval fails or when the model misinterprets retrieved content. Moreover, the presence of sources does not guarantee that the response is directly based on them, especially when the model blends internal knowledge with retrieved facts. Existing explainability tools offer limited support here;

they do not explain hallucinations in a structured or reliable way. Bridging this gap requires a deeper understanding of how these models reason—and how they fail to do so[3,40].

Use case: hallucinations in Decision Support Systems

Decision Support Systems (DSS) are using LLMs more and more to help professionals with difficult, knowledge-and reasoning-intensive tasks. These systems frequently function in crucial fields where precise, context-sensitive data is crucial, including public administration, healthcare, law, and finance. LLMs support DSS by performing tasks such as answering domain-specific questions, summarizing documents, producing structured reports, finding and combining information from large knowledge bases, and explaining complex concepts in plain language. Their ability to understand and produce human-like language makes them especially well-suited for facilitating communication between users and structured or unstructured data sources[5]. Integrating LLMs into DSS aims to improve access to expert-level knowledge, reduce human error, and increase efficiency. However, hallucinations pose serious risks, particularly when outputs appear authoritative but contain factual or logical errors[16,41].

Risks in medical and legal applications

Consider a medical DSS that supports clinicians in diagnosis or treatment planning. If the LLM fabricates a clinical guideline or invents statistics about drug interactions, it can mislead practitioners and jeopardize patient safety. This risk is heightened by the model's tendency to express false information with confidence and fluency, making it appear credible. Clinicians may lack the time or expertise to verify every suggestion, particularly under time pressure or in resource-limited settings[42].

Similarly, in a legal DSS, an LLM might invent a legal precedent or misquote a law. Even if the overall argument is coherent, citing a non-existent case or misinterpreting a statute can result in flawed legal reasoning. This can undermine legal research and courtroom arguments, especially when non-experts rely on the system as a primary information source. In both medical and legal contexts, such hallucinations not only mislead users but also raise ethical and legal risks for deploying institutions[43].

The explainability gap in DSS outputs

These failures are especially problematic because the reasoning behind LLM outputs is often opaque. Most models do not reveal their internal logic or the sources that shaped their responses. Without a way to trace or challenge the reasoning process, users are left with outputs that may appear correct but cannot be reliably trusted. Even RAGenhanced systems, while providing supporting documents, often fail to clarify how the final answer is derived from the retrieved content[44,45].

Significance of the use case

This use case emphasizes the urgent need for explainability frameworks that go beyond defending an output's appearance, particularly in clinical diagnosis and legal support. The ability to evaluate whether a model's response adheres to domain constraints, is supported by trustworthy data, and follows sound reasoning patterns is a crucial prerequisite[40]. Without such transparency, LLM hallucinations become more than technical flaws— they pose real risks in high-stakes decisions. By anchoring our theoretical discussion in a real-world example, we show that hallucinations are not merely a research concern, but a critical barrier to the safe, responsible deployment of LLMs in decision-critical systems. Addressing their explainability is therefore essential[46].

Comparative analysis of explainability methods

We present a comparative analysis of popular explainability techniques applied to hallucinated outputs in LLMs in order to investigate the shortcomings of existing explainability tools. This systematic comparison looks at how well each approach works in three important areas: identifying hallucinations, elucidating their causes, and providing local or global explainability. We contrast six widely used explainability techniques: Training data attribution, prompt-based justifications, saliency maps, attention maps, SHAP, and LIME[6,31]. The comparison is based on four key criteria:

- **Hallucination detection**: Can the method identify when an output is misleading or false?
- **Explanation of causes**: Can the method explain why the hallucination occurred?
- **Explainability scope**: Does the method offer insights at the global model level (i.e., overall behavior) or at the local level of individual outputs? (See Section **??** for definitions.)
- **Method type**: Does the method generate explanations after the output is produced (post-hoc), or are the explanations built into how the model makes decisions (intrinsic)?

These findings are compiled in the table below:

The values in this table are based on synthesis of findings reported in prior surveys and tool documentation[5,6,47]. In some cases, "Partial" or "Sometimes" reflects inconsistent performance or dependency on implementation choices.

The analysis reveals that most existing methods struggle to provide causal or trainingaware explanations for hallucinations. Below, we provide method-specific justifications for the values in Table 1, focusing in particular on the dimensions of hallucination detection and explanation of causes.

SHAP and LIME: These methods assign feature importance scores based on the relative contributions of each input token to the model's output. Tokens with little to no semantic relevance to the output may be given weight by SHAP or LIME when working with hallucinated outputs. An *indirect signal* that the output is unreliable—that is, not consistently connected to the input or external facts—can be this disparity between

Table 1. Comparison of explainability methods for hallucination analysis in LLMs.

Method	Detects Hallucination	Explains Cause	Scope	Type
SHAP	Partial	No	Local	Post-hoc
LIME	Partial	No	Local	Post-hoc
Attention maps	No	No	Local	Intrinsic
Saliency maps	No	No	Local	Intrinsic
Prompt-based justifications	Sometimes	Partially	Local	Post-hoc
Training data attribution	Yes	Yes	Global	Intrinsic

semantic grounding and importance. However, these methods don't look at the output's factual consistency or reasoning process, nor do they evaluate whether outside knowledge was misused. Their "Partial" detection and "No" explanation capabilities are thus a result[48,49].

Attention and Saliency Maps: Saliency maps pinpoint input regions with significant gradient influence, while attention visualizations draw attention to token-to-token weights in transformer layers. Both are descriptive tools that show what inputs "mattered" during generation or where the model "looked." They do not, however, provide a way to assess whether the output is grounded or factually accurate. They offer neither detection nor explanation because they are unable to discern between a coherent hallucination and an honest response. They are rated as "No" in both criteria due to these factors[50,51].

Prompt-based Justifications: By asking the LLM to defend its response in natural language, these techniques elicit explanations from the model. Even though they could be helpful for transparency, these explanations frequently highlight the model's linguistic fluency rather than its accuracy. To further mislead users, an LLM might, for instance, produce a confident but inaccurate explanation for a hallucinated response. Because they may occasionally reveal inconsistencies, their detection capability is rated as "Sometimes" and their ability to explain hallucinations as "Partial" at best[33,52].

Training Data Attribution: These approaches use techniques like nearest-neighbor search in embedding space to link model outputs to significant training examples. The answer may be deemed hallucinatory or generalized beyond the training data if no comparable examples are found or if the traced examples conflict with the output. Furthermore, these techniques aid in determining whether a hallucination is the consequence of bias, memorization, or incomplete data. They are thus the only approach that has been given a "Yes" rating for both detection and causal explanation[53,54].

In this case, terms such as "Partial" indicate that the method offers shaky, evasive, or inconsistent support for the pertinent criterion. Here, "untrustworthy" refers to outputs whose token-level attributions or explanations don't match the factual accuracy or logical consistency of the generated text. In other words, while some methods may suggest that "something is amiss," they rarely provide the tools to identify what or why. This comparison highlights the difference between surface-level interpretability and more thorough, logical explanation. Closing this gap may require new approaches that consider the internal reasoning of models, which we explore in the next section.

Towards more trustworthy explanations

It is necessary to address deeper sources of failure like reasoning gaps, training data limitations, and uncalibrated confidence in order to improve the credibility of LLM explanations. This goes beyond surface-level feature attribution, which refers to techniques that highlight which input features influenced the output. Key tactics for improving explanation reliability in practical applications are described in this section[55].

Hybrid reasoning and data transparency

Deterministic reasoning can improve traceability and constrain outputs in stochastic models. Domain-specific logic violations can be avoided and the reasoning process of the model can be made clearer with hybrid approaches that incorporate rules, symbolic logic, or knowledge bases. It is equally important to be transparent about the sources of training data[3]. A lot of hallucinations are caused by biased or incomplete datasets, but users are frequently ignorant of these flaws. Attribution techniques that connect outputs to significant training examples can offer important context when model knowledge is unclear[56].

Uncertainty awareness and conservative generation

Another essential component of credible explainability is confidence calibration. Users tend to overtrust hallucinated content because LLMs frequently express uncertainty with excessive fluency. Methods like probabilistic scoring and token-level confidence estimates can assist in bringing perceived reliability and actual model certainty into alignment. Further reducing speculative outputs can be achieved by prompting strategies, such as telling the model to reply with "I don't know" when

unsure[57]. Furthermore, when the model lacks adequate knowledge, fine-tuning using techniques like Reinforcement Learning from Human Feedback (RLHF) encourages honesty and more conservative behavior[55,57].

Conclusion

One of the most persistent issues with LLMs is hallucinations, especially in domains where accurate facts and clear reasoning are crucial. Despite their clear communication, LLMs often generate content that is erroneous, nonsensical, or unproven. The explainability techniques currently in use, like attention maps, SHAP, and LIME, rarely reveal the underlying logic that leads to these errors.

In this study, we reframed hallucinations as indicators of reasoning errors rather than as distinct output anomalies. We examined contributing elements like unclear prompts and skewed training data and made a distinction between reasoning-based and factual hallucinations. Through a comparative analysis, we evaluated the strengths and weaknesses of the existing interpretability tools in addressing these problems.

We argued that in order to improve explainability, a deeper understanding of LLM reasoning is required. Deterministic methods provide logical structure and traceability, whereas stochastic models offer flexibility. Combining both perspectives in a hybrid reasoning-aware approach can increase the reliability of model behavior. By using practical strategies like training data attribution, confidence calibration, and encouraging models to abstain when uncertain, reliability and safety can be further enhanced. Fundamentally, explainability needs to go beyond cursory justifications. An output that contains errors should provide an explanation of how and why it was created. Explanations that take uncertainty and internal reasoning into account are essential for the responsible use of LLMs in decision-making situations.

Data and software availability

No original datasets or computational models were generated or used in this study. The analysis and discussion are based solely on information available in the referenced literature.

References

1. Mikhailov DI: **Optimizing national security strategies through LLM-driven artificial intelligence integration.** 2023.
Publisher Full Text
2. Williamson SM, Prybutok V: **The era of artificial intelligence deception: unraveling the complexities of false realities and emerging threats of misinformation.** *Information.* 2024; **15**: 299.
Publisher Full Text
3. Rumiantsau M, Vertsel A, Hrytsuk I, *et al.*: **Beyond fine-tuning: effective strategies for mitigating hallucinations in large language models for data analytics.** 2024.
Publisher Full Text
4. Koenecke A, Choi ASG, Mei KX, *et al.*: **Careless whisper: speech-to-text hallucination harms.** *2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024.* 2024; 1672–1681.
Publisher Full Text
5. Gomes C, Natraj L, Liu S, *et al.*: **A survey of explainable AI and proposal for a discipline of explanation engineering.** *arXiv preprint arXiv: 2306.01750.* 2023.
Publisher Full Text
6. Mumuni F, Mumuni A: **Explainable artificial intelligence (xai): from inherent explainability to large language models.** 2025.
Publisher Full Text
7. Petrovic D: **Advanced interpretability techniques for tracing LLM activations.** Dejan.ai blog, published Mar 31, 2025. Covers activation logging, causal tracing with activation patching, logit lens and attention/neuron-level interpretability techniques. 2025.
Reference Source
8. Zaman K, Srivastava S: **A causal lens for evaluating faithfulness metrics.** arXiv preprint arXiv: 2502.18848. 2025.
Publisher Full Text
9. Tonmoy SMTI, Zaman SMM, Jain V, *et al.*: **A comprehensive survey of hallucination mitigation techniques in large language models.** 2024.
Publisher Full Text
10. Banerjee S, Agarwal A, Singla S: **Llms will always hallucinate, and we need to live with this.** 2024.
Publisher Full Text
11. Barros S: **I think, therefore i hallucinate: minds, machines, and the art of being wrong.** 2025.
Publisher Full Text
12. Rodu J, Baiocchi M: **When black box algorithms are (not) appropriate.** *Obs Stud.* 2023; **9**(2): 79–101.
Publisher Full Text
13. Kabra S, Tech V, Jha A, *et al.*: **Reasoning towards fairness: mitigating bias in language models through reasoning-guided fine-tuning.** 2025.
Publisher Full Text
14. Berberette E, Hutchins J, Sadovnik A: **Redefining "hallucination" in LLMs: towards a psychology-informed framework for mitigating misinformation.** 2024.
Publisher Full Text
15. Zhao H, Chen H, Yang F, *et al.*: **Explainability for large language models: a survey.** *ACM Trans Intell Syst Technol.* 2024; **15**(2): 1–38.
Publisher Full Text
16. Ye H, Liu T, Zhang A, *et al.*: **Cognitive mirage: a review of hallucinations in large language models.** 2023.
Publisher Full Text
17. Sahoo P, Meharia P, Ghosh A, *et al.*: **A comprehensive survey of hallucination in large language, image, video and audio foundation models.** 2024.
Publisher Full Text
18. Maree C, Modal JE, Omlin CW: **Towards responsible AI for financial transactions.** *2020 IEEE Symposium Series on Computational Intelligence, SSCI.* 2020; 16–21.
Publisher Full Text
19. Hall P: **On the art and science of machine learning explanations.** 2018.
Publisher Full Text
20. Santana J: **Deliberate, analytical, and rule-based reasoning.** LinkedIn article, published 2 June 2025.
Reference Source
21. Hopgood AA: **The state of artificial intelligence.** *Advances in Computers.* 2005; **65**: 1–75.
Publisher Full Text
22. Besta M, Barth J, Schreiber E, *et al.*: **Reasoning language models: a blueprint.** 2025.
Publisher Full Text
23. Peng B, Chen K, Li M, *et al.*: **Securing large language models: addressing bias, misinformation, and prompt attacks.** 2024.
Publisher Full Text
24. Jannadi KA: **Enhancing the reliability of Large Language Models and mitigating hallucinations and inaccuracies.** *Medium article.* 2023.
Reference Source

25. Janumpally R, Nanua S, Ngo A, *et al.*: **Generative artificial intelligence in graduate medical education.** *Front Med (Lausanne).* 2025; **11**: 1525604.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

26. Feldman P, Foulds JR, Pan S: **Ragged edges: the double-edged sword of retrievalaugmented chatbots.** 2024.
Publisher Full Text

27. Kalai AT, Ai O, Francisco S, *et al.*: **Calibrated language models must hallucinate.** *Proceedings of the Annual ACM Symposium on Theory of Computing.* 2024; 160–171.
Publisher Full Text

28. Kim H, Bibi A, Torr P, *et al.*: **Detecting LLM hallucination through layer-wise information deficiency: analysis of unanswerable questions and ambiguous prompts.** 2024.
Publisher Full Text

29. Lee K, Firat O, Agarwal A, *et al.*: **Hallucinations in neural machine translation.** In: *International Conference on Learning Representations (ICLR).* under review/double-blind. 2019.
Reference Source

30. Hatem R, Simmons B, Thornton JE: **A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks.** *Cureus.* 2023; **15**(9): e44720.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

31. Huang L, Yu W, Ma W, *et al.*: **A survey on hallucination in Large Language Models: principles, taxonomy, challenges, and open questions.** *58 ACM Trans Inform Syst.* 2023; **1**.
Publisher Full Text

32. Das D, Banerjee S, Chernova S: **Explainable AI for system failures: generating explanations that improve human assistance in fault recovery**. 2020.
Publisher Full Text

33. Muhammad D, Bendechache M: **Unveiling the black box: a systematic review of explainable artificial intelligence in medical image analysis.** *Comput Struct Biotechnol J.* 2024; **24**: 542–560.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

34. Wan A, Dunlap L, Ho D, *et al.*: **Nbdt: Neural-backed decision trees.** *ICLR 2021 - 9th International Conference on Learning Representations*. 2020.
Publisher Full Text

35. Jesson A, Beltran-Velez N, Chu Q, *et al.*: **Estimating the hallucination rate of generative AI.** arXiv preprint arXiv: 2406.07457, submitted 11 Jun 2024, June, 2024.
Publisher Full Text

36. Saba WS: **Stochastic LLMs *do not* understand language: towards symbolic, explainable and ontologically based LLMs.** *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) LNCS, 2023; **14320**: 3–19.
Publisher Full Text

37. Wang L, Yang N, Huang X, *et al.*: **Improving text embeddings with Large Language Models.** *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 2024; **1**: 11897–11916.
Publisher Full Text

38. Verma S: **Contextual compression in retrieval-augmented generation for Large Language Models: a survey.** arXiv preprint arXiv: 2409.13385, last revised 2 Oct 2024, Sep, 2024.
Publisher Full Text

39. Wu K, Wu E, Wei K, *et al.*: **An automated framework for assessing how well LLMs cite relevant medical references.** *Nat Commun.* 2025; **16**(1): 3615.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

40. Rudin C: **Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.** *Nat Mach Intell.* 2019; **1**(5): 206–215.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

41. Pham DK, Vo BQ: **Towards reliable medical question answering: techniques and challenges in mitigating hallucinations in language models**. 2024.
Publisher Full Text

42. Bridge P, Fielding A, Rowntree P, *et al.*: **Intraobserver variability: should we worry?** *J Med Imaging Radiat Sci.* 2016; **47**(3): 217–220.
PubMed Abstract | **Publisher Full Text**

43. Zhui L, Fenghe L, Xuehu W, *et al.*: **Ethical considerations and fundamental principles of Large Language Models in medical education: viewpoint.** *J Med Internet Res.* 2024; **26**: e60083.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

44. Vladika J, Matthes F: **On the influence of context size and model choice in retrievalaugmented generation systems**. 2025.
Publisher Full Text

45. Wong L, Ali A, Xiong R, *et al.*: **Retrieval-augmented systems can be dangerous medical communicators**. 2025.
Publisher Full Text

46. Dahl M, Magesh V, Suzgun M, *et al.*: **Large legal fictions: profiling legal hallucinations in Large Language Models.** *J Legal Anal.* 2024; **16**(1): 64–93.
Publisher Full Text

47. Molnar C: **Interpretable machine learning.** Lulu.com, online book on interpretable ML, widely cited for explainability frameworks. 2020.
Reference Source

48. Lundberg SM, Lee SI: **A unified approach to interpreting model predictions.** In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017; 4765–4774.
Reference Source

49. Ribeiro MT, Singh S, Guestrin C: **"why should i trust you?": explaining the predictions of any classifier.** In: *Proceedings of the 22nd ACM SIGKDD*. 2016; 1135–1144.
Publisher Full Text

50. Wiegreffe S, Pinter Y: **Attention is not explanation**. arXiv preprint arXiv: 1908.04626, 2019.
Publisher Full Text

51. Adebayo J, Gilmer J, Muelly M, *et al.*: **Sanity checks for saliency maps.** In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018; 9505–9515.
Reference Source

52. Lampinen ENM, Clark K, Goodman ND, *et al.*: **Can language models learn from explanations? A dataset and baseline results for answering, explaining and verifying answers.** arXiv preprint arXiv: 2204.02329, 2022.
Publisher Full Text

53. Carlini N, Tramer F, Wallace E, *et al.*: **Extracting training data from diffusion models.** arXiv preprint arXiv: 2301.13188, 2023.
Publisher Full Text

54. Zhou Y, Booth S, Ribeiro MT, *et al.*: **Do feature attribution methods correctly attribute features?** *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*. 2021; **36**(9): 9623–9633.
Publisher Full Text

55. Wu X, Zhao H, Zhu Y, *et al.*: **Usable xai: 10 strategies towards exploiting explainability in the LLM era**. 2024.
Publisher Full Text

56. Sun Y, Gai Y, Chen L, *et al.*: **Why and how LLMs hallucinate: connecting the dots with subsequence associations**. 2025.
Publisher Full Text

57. Dam H, Knochelmann J, Joseph V, *et al.*: **Derailing non-answers via logit suppression at output subspace boundaries in RLHF-aligned language models**. 2025.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✗ ✗ ?

Version 1

Reviewer Report 31 December 2025

https://doi.org/10.21956/openreseurope.22547.r64532

© **2025 Yaprak B.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Burak Yaprak
İstanbul Ticaret University,, İstanbul, Turkey

Intro
This review article addresses hallucinations in Large Language Models (LLMs), a critical challenge that undermines the reliability of AI systems, particularly in high-risk decision support settings such as medical diagnosis and legal analysis. The authors aim to reframe hallucinations not as mere output anomalies, but as reasoning failures, and to assess the limitations of existing Explainable AI (XAI) approaches (e.g., SHAP, LIME, attention mechanisms) in uncovering the causes of such failures.

Strengths
Critical Problem: The manuscript tackles a timely and highly relevant issue by focusing on the root causes of hallucinations in LLMs, a central concern in the broader discussion on trustworthy and responsible AI.

Framework: The proposal to conceptualize hallucinations as reasoning failures, and to distinguish between factual and reasoning-based hallucinations, adds meaningful theoretical depth. This distinction could serve as a useful foundation for future work on hallucination detection and mitigation.

Readability: Despite addressing a technically complex topic, the manuscript is written in a clear, accessible, and well-structured manner. The introduction and conceptual explanations are easy to follow, even for readers who are not deeply specialized in the area.

Concerns
Lack of Methodological Transparency (Critical Issue): The paper's central contribution, the comparative analysis of XAI methods for hallucination analysis, particularly as presented in Table 1, is not supported by either an empirical study or a clearly defined systematic review methodology (e.g., PRISMA). The qualitative assessments used in the table (e.g., "Partial," "No," "Sometimes") are presented without an explicit and reproducible evaluation framework. As a result, the scientific validity and reproducibility of the core findings are difficult to assess.

RAG: RAG is currently one of the most widely used and practically effective approaches for mitigating hallucinations in LLMs. Importantly, RAG also provides a form of explanation by explicitly linking generated outputs to retrieved source documents. However, RAG is not included in the main comparison table and is only briefly discussed in the text. This represents a significant omission in the context of contemporary hallucination analysis and explainability.

Classification Inconsistencies: The manuscript's categorization of certain XAI methods is debatable. For instance, labeling attention maps as "intrinsic" methods conflicts with the prevailing view in the XAI literature, where such techniques are typically regarded as post-hoc attribution methods. These terminological and classification issues weaken the conceptual foundations of the paper.

Is the topic of the review discussed comprehensively in the context of the current literature?
Partly

Are all factual statements correct and adequately supported by citations?
No

Is the review written in accessible language?
Yes

Are the conclusions drawn appropriate in the context of the current research literature?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research focuses on the use of AI systems in marketing applications, with particular attention to hallucinations and their implications in real-world use cases. I have recently published a paper addressing this topic, and my expertise lies primarily in understanding hallucinations from an applied and literature-based perspective rather than from a deeply technical or model-centric standpoint.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 23 December 2025

https://doi.org/10.21956/openreseurope.22547.r64535

© 2025 Graells A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Albert Sánchez Graells
Bristol University, Bristol, UK

The article focuses on large Language Model (LLM)'s hallucinations as a key technical challenge and a driver of risk for the adoption of generative AI in the context of high-stakes decision-making--which is used as an example of deployment to highlight the risks arising from hallucinations and to collate some anecdotal evidence on their materialisation in the context of healthcare and legal decision-making. This is an important and current issue with major governance implications.

The article focuses in particular on the analysis of hallucinations as an instance of defective reasoning leading, potentially, to factual hallucinations and reasoning-based hallucinations. It then seeks to link these issues with AI explainability challenges and critically assesses some XAI methods to discard them and to posit their insufficiency or inadequacy in relation to hallucinations.

In my view, the article does not clearly make the connection between hallucinations and explainability and it fails to explain why thinking about the challenge of hallucinations from this perspective could be helpful. I struggle to see the differentiation between factual and reasoning-based hallucinations, given that there is significant overlap between both of them (as the article acknowledges). More importantly, I do not think explainability is a key issue in this context. Explainability is crucially important in the context of symbolic AI because its use in the high-stakes contexts explored in the article requires the giving of reasons (on which there is a rich literature) to enable different expert and governance checks and balances to be deployed. However, I am not sure what the role of explainability in LLM use would be. If, to use the example in the article, an LLM outputs that the Tour Eiffel is in Berlin, it does not really matter why this was outputted (other than to the developers of the LLM, perhaps). Moreover, in most cases, what the article seems to be concerned with (in its discussion and problematisation) is with hallucination detection, rather than hallucination origination. In that regard, the connection between these issues needs to be revisited and, if there is a case for focused analysis from an XAI perspective, that case needs to be made clearly and persuasively.

Is the topic of the review discussed comprehensively in the context of the current literature?
Partly

Are all factual statements correct and adequately supported by citations?
Partly

Is the review written in accessible language?
Partly

Are the conclusions drawn appropriate in the context of the current research literature?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Digital governance, public sector AI adoption, AI procurement

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 04 August 2025

https://doi.org/10.21956/openreseurope.22547.r57102

© 2025 Leiser F. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Florian Leiser
Technical University Munich, Munich, Germany

==== General Remarks ====
This review paper investigates a comparison of different explainability methods for hallucination analysis in LLMs. By summarizing existing literature, the authors claim to address four key objectives: (1) the summarization of current findings in LLM hallucinations, (2) an evaluation of interpretability techniques “through the lens of reasoning”, (3) introducing the intersection of deterministic and stochastic reasoning in explainability, (4) comparing explainability in a DSS use case.
I am writing this review from the perspective of a researcher in information systems with a focus on human-computer interaction. Specifically, I am investigating (user-centered) hallucination mitigation in LLMs with some basic interest in explainable AI (XAI). I already worked on scoping reviews, but I am not familiar with the deterministic/stochastic reasoning paradigms the authors use in their studies.
==== Major Strengths ====
The paper investigates the timely and relevant topic of hallucination mitigation in LLMs by using explainability methods. Their perspective can spark discussion by bringing reasoning perspectives into the discussion and highlighting different perspectives on LLM reasoning. The paper is well-written and is accessible in its language.
==== Major Issues ====
I do have some concerns regarding this study especially to the methodology and the connection to established literature which are presented below.

1. **Methodology:** My major issue is on the reproducibility of the analysis. The authors do not state how they came to the literature they are citing, how they abstracted their insights, or how they came up with their comparison of explainability methods. For the selection of manuscripts, I would have benefited from established guidelines (such as PRISMA)[1] to report the literature. For the comparison of XAI methods, I would have needed additional information why those four criteria were selected and why the authors focused on those explainability methods. I was also wondering how the comparison of the XAI methods was achieved. Was any empirical evidence (literature or simulation-wise) involved, or does it purely reflect the opinion of the authors?
2. **Motivation**: I am struggling with some aspects of the motivation. Since the authors

themselves claim that hallucinations are part of LLMs and can never be completely mitigated, the manuscript could further strengthen the need for explaining hallucinations. The authors highlight the distinction between factual and reasoning-based hallucinations while later introducing hallucinations by training data noise, prompt design, or model probabilities. I would need stronger arguments on why the authors focus on certain types of hallucinations instead of others, for example presented in Ji et al. (2023)[2].
Similarly, I was wondering about selecting DSS as an example? What is special about DSS (or the DSS-LLM-combination) where hallucinations are more important?

3. **Imprecise concept definitions**: Throughout the manuscript, terms and concepts are used inconsistently, making it challenging to understand the claims of the authors. For example, transparency, explainability, and interpretability are used interchangeably although there are established differences in their notion (e.g., Arrieta et al. (2020)[3]). According to Arrieta et al. (2020)[3], attention maps and saliency maps are both considered as post-hoc explainability methods contrasting the statements of the authors of this review. Finally, XAI explains the output of the model. The authors, however, try to contextualize XAI on hallucinations (i.e., wrong predictions). Using XAI to explain why a output is wrong needs to have a second-level analysis to it.
4. **Presentation:** When reading the manuscript, I felt lost between the different sections. As suggested in the ORE guidelines, the recommended structure is Introduction, Discussion, Conclusions. The authors provide their own structure by introducing a background section (which I see as beneficial) and subsections on their objectives. Using this structure, they jump between the different concepts (i.e., hallucinations, explainability). For example, in the background they “identify two main categories” for hallucinations which are only introduced later in another chapter. Streamlining these sections for a coherent reading flow would help me in understanding the different concepts. In addition, there are minor inconsistencies in the usage of abbreviations and formulations, impairing the overall paper presentation.

==== Minor Issues ====

- ORE suggests that “Reviews should be based on peer-reviewed literature, and should not include new research, data or propose new hypotheses.” Throughout the references, about half of the manuscripts are not peer-reviewed references (e.g., from arXiv, SSRN, or LinkedIn), although peer-reviewed paper are available (see examplary sources attached)
- I don’t fully buy the argument why hallucinations should be reasoning errors. LLMs are probability-based models and reason to the best of their knowledge resulting in a wrong prediction. Contextualizing this to GenAI tools in image creation would mean that the reasoning process from input to image would be false, and not the generated image itself. I don’t see (or understand) the arguments of the authors why this perspective would help us address the mitigation of hallucinations and encourage the authors to revise and strengthen their argumentation here.
- Using figures and tables would help me understand how the authors see the connection between the concepts of explainability, hallucinations, and reasoning.
- In the comparison of the explanation methods in Table 1, I was wondering how a global explanation method can detect hallucinations that occur on an individual prediction? I would benefit from an elaboration on that.
- To allow for additional space to incorporate the comments, I would suggest to drop the meta-comments describing what will happen in this section (especially in the background)
- Finally in the discussion, the authors propose novel ideas on how hallucinations might be mitigated using other explanations strategies. I would encourage the authors to introduce them early on, since I believe this would help strengthen the relevance of this review.

References
1. Page M, McKenzie J, Bossuyt P, Boutron I, et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021. Publisher Full Text
2. Ji Z, Lee N, Frieske R, Yu T, et al.: Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*. 2023; **55** (12): 1-38 Publisher Full Text
3. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. 2020; **58**: 82-115 Publisher Full Text

Is the topic of the review discussed comprehensively in the context of the current literature?
Partly

Are all factual statements correct and adequately supported by citations?
No

Is the review written in accessible language?
Yes

Are the conclusions drawn appropriate in the context of the current research literature?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Information Systems, Human-Computer-Interaction, Hallucination Mitigation, Explainable AI

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.