AI & SOCIETY
https://doi.org/10.1007/s00146-025-02757-1

OPEN FORUM

'AI lost the prompt!' Replacing 'AI hallucination' to distinguish between mere errors and irregularities

José María Ariso[1] · **Peter Bannister**[1]

Received: 24 June 2025 / Accepted: 16 November 2025

Abstract
One of the principal areas of current AI research concerns what are termed "hallucinations". Whilst hundreds of different definitions and classifications of "AI hallucination" have been published, none have yet considered the distinction between errors and irregularities in Wittgenstein's sense. This article provides a straightforward explanation of this distinction, illustrated through examples of AI outputs drawn from various publications. We then examine the terms proposed as alternatives to "hallucination" and highlight both their strengths and weaknesses. Drawing upon this analysis, we establish criteria for proposing alternative terms that encompass both errors and irregularities in Wittgenstein's sense. Our aim is not to definitively resolve the ongoing debate surrounding suitable replacements for "AI hallucination", but rather to provide a comprehensive overview of the characteristics and nuances that this distinction brings to the discussion. For unlike errors, irregularities prove entirely incomprehensible to users, owing to the grammatical gap created when fundamental certainties that underpin meaningful language use are violated. Given that irregularities prove incomprehensible, as they violate the certainties underlying meaningful language use, the most trustworthy AI systems may ultimately be those that recognise their own epistemic boundaries rather than those that produce seemingly perfect outputs.

Keywords Hallucination · AI systems · Human–AI interaction · Error · Certainty · Hinge

1 Introduction

Popular media, commercial messages and even scientific articles often fall into the trap of representing AI systems as shiny humanoid robots or free-floating electronic brains (Bannister and Carver 2024). These images wrongly lead people to think that robots, much like ourselves, are able to act autonomously and solve problems independently. Most importantly, such images may even lead people to think that AI systems can reason in ways superior to them (Steen et al. 2024). In line with this, the CEO of Tesla Elon Musk predicted in 2024 that AI will surpass human intelligence by 2026 (The Guardian 2024), not forgetting that AI systems surpassed humans some years ago at games like six-player poker (Blair and Saffidine 2019) and Go (BBC 2017). Some scholars contend that the predictive accuracy achieved by AI systems in the medical field grants these technologies a form of epistemic expertise and authority over human experts (Bjerring and Busch 2021; Grote and Berens 2020). Indeed, a team of licenced health care professionals preferred ChatGPT's responses to patients' questions over physician responses, and even rated AI-generated response significantly higher in terms of both quality and empathy (Ayers et al. 2023). As if this were not enough, nearly three thousand researchers predicted in 2023 that there is a 10% chance of AI outperforming humans in all tasks by 2027, and this rises to 50% by 2047 (Grace et al. 2024). This suggests that AI progress is faster than expected, for, when the same survey was carried out in 2022, they forecasted a 10% chance of high-level machine intelligence arriving by 2029, and a 50% chance by 2060 (Grace et al. 2022), so only 1 year later, experts predicted that this could happen 13 years sooner.

The combination of AI's immediacy, accuracy and comprehensive information provision with media hype may lead many to perceive AI technologies as magical (Leaver and Srdarov 2023). It should, therefore, come as no surprise that

✉ José María Ariso
josemaria.ariso@unir.net

✉ Peter Bannister
peter.bannister@unir.net

[1] Universidad Internacional de La Rioja, Logroño, Spain

Published online: 24 November 2025

one of the myths about AI is that it can solve any problem,[1] with the implication that AI would commit no errors in doing so. Many users are, thus, expected to believe that they are in an inferior position to AI when interacting with it inasmuch as it is an infallible or unerring tool. However, AI makes and will continue to make mistakes (Bannister in press). Given that the phenomenon whereby AI provides distorted information—usually called "hallucination" (Sun et al. 2024)—has been regarded as an inherent feature of Open AI's ChatGPT, as well as of other AI systems, its complete eradication without compromising the high-quality performance of ChatGPT is almost impossible (Lee 2023). This claim allows us to highlight two aspects that form the backbone of our paper.

On the one hand, the use of the term "hallucination" to refer to AI outputs has enjoyed rapid diffusion at both popular and specialist levels, which seemingly entails that it could have happened only if its meaning were clear enough for all users. However, the bibliographic analysis carried out by Maleki et al. (2024) includes 333 distinct definitions of "AI hallucination", a number that has continued to grow without any consensus being reached. This is striking if we take into account that the term "hallucination" has been used, notwithstanding some exceptions (e.g. Izadi and Forouzanfar 2024), to designate something apparently as homogeneous as the practical totality of cases in which AI provides distorted information. However, the problem in finding a common and agreed-upon denomination is the great diversity of such cases. By way of example, amongst these cases are mixed indiscriminately examples of what the philosopher Ludwig Wittgenstein (1997) considered as cases of "error" and "irregularity". This difference, which has gone unnoticed in the literature thus far, is relevant insofar as it seems to have been pointed out by ChatGPT itself when recognising that it "can sometimes make mistakes or generate illogical responses" (Borji 2023, p. 6).

Moreover, this distinction is of utmost importance for user trust in AI systems. For AI-made errors will make the technology appear less reliable to the user; but if AI failures correspond to irregularities, as we will clarify in the following section, a grammatical gap will occur between the AI system and the user (Ariso 2022a), who will perceive said system as an interlocutor they cannot truly understand. In other words, although AI cannot avoid errors, if it at least avoided irregularities, it would be perceived as much more reliable and, what is more important, as more relatable by users. Therefore, in this article, we shall begin by offering a brief exposition illustrated with simple examples of the distinction that Wittgenstein established between "error" and "irregularity", a distinction that has not been taken into account at all when establishing classifications of hallucinations in AI. Subsequently, we analyse various cases collected in the literature to distinguish which constitute examples of error or irregularity. Based on this foundation, we explain why AI would be perceived as less relatable by users when it provides outputs that can be regarded as instances of irregularity or madness in Wittgenstein's sense.

On the other hand, recent literature presents several reasons for which the term "hallucination" is considered inadequate for referring to AI errors which we will address in greater depth subsequently. We further the scope of this discussion by proposing some alternatives that allow us to distinguish between errors and irregularities. To that end, we shall analyse the terms that have been proposed as alternatives, pointing out their pros and cons. To choose the terms we propose as substitutes for "hallucination", we shall elaborate a list of criteria, taking into account the reasons put forward for rejecting the term "hallucination", the analysis of alternative terms, and criteria such as simplicity, the absence of negative connotations or AI's lack of intentionality, amongst others. This search for alternative terms has as its fundamental objective not so much in finding exact concepts to substitute the term "hallucination" but rather illustrate in greater detail the distinction between "error" and "irregularity" to facilitate better understanding of this difference. For it can be of great help in understanding what type of errors AI systems should avoid to be more similar to humans in their performance, which would generate greater trust amongst AI users.

2 Wittgenstein's distinction between errors and irregularities

Let us suppose that someone is asked how many people remained in the small lift from which they have just exited. Let us also suppose that there were four people. If the subject in question were to respond that three or five people, we would not hesitate to say that they had made a mistake or an error. But if they were to say that 17,326,579 people remained in the lift, we would think that this is such an out-of-place response that it cannot just be considered as an error. Just as in the first case, one could justify this by saying that three or five people remained because they did not see a person who was hidden behind another, or because they counted one person twice, in the second case, we would not know what to even think as to the reasons that led the subject to respond that so many people were still in the lift. As Wittgenstein (1997, §673) acknowledged, the boundary between both types of cases would not be clear. For where would the boundary be between the erroneous response and the anomalous one? Would it be in saying that there were 20, 50, or 100 people? However, it is evident that there are

[1] https://www.aimyths.org/.

responses that cannot be considered as mere errors. Thus, although Descartes (1986, p. 38) claimed to know by experience that he was "prone to countless errors", there are many errors that we cannot make.

Other examples of errors that we cannot make would consist in affirming that we are not alive, that our family and friends do not have minds, that physical objects do not exist, or what the meaning of the most basic words of our mother tongue is, amongst many others. What this type of error that cannot be considered as such has in common is that it violates what Wittgenstein (1997) called "hinge" or "certainty". These certainties constitute implicit assumptions in everything we say and do, but they are not grounded in specific reasons. For even if we could provide many grounds for supporting some certainty, none would be "as certain as the very thing they were supposed to be grounds for" (Wittgenstein 1997, §307). By way of example, someone could say that they have a body because they saw it reflected in a mirror. However, if they did not see their body reflected when looking in a mirror, that would not mean that they had lost it, but rather that they had a perceptual disorder. Thus, although that reason might seem to indicate that they did not have a body, both all of us and the subject in question would maintain the certainty that they do indeed have a body. From this, it follows that seeing one's own body reflected in a mirror does not constitute a reason that can sustain the certainty that the aforementioned subject has a body even when it might seem to indicate so.

Another important characteristic of certainties is that they are neither acquired (Ariso 2022b) nor lost (Ariso 2025a) willingly. This can be easily verified, for example by trying to acquire the certainty that all red-haired people are extraterrestrials, or by trying to lose the certainty that walls can be passed through as if they were mere mirages. However, these examples should not lead us to think that certainties are mental states. As said above, they are implicit assumptions in everything we say and do, so they can be shown with very different mental states. Indeed, we constantly show our certainty of having a body regardless of whether we are very happy or very sad. What must be taken into account to discern whether something is a certainty is that the possibility of doubt or error about it is logically excluded (Wittgenstein 1997, §194). Yet such exclusion is not located "in the traditional and context-independent logic, which is valid in all possible worlds, but in the particular logic that emanates from our linguistic practices" (Ariso 2020, pp. 659–660). Bearing this in mind, Coliva (2010) indicates that a doubt, to be truly considered as such and not fall into what Wittgenstein (1997, §249) calls "a false picture of doubt", must fulfil these four criteria. First, doubts manifest themselves only when the context allows for them, and consequently, this must have practical implications. Second, doubts must be based on grounds which allow us to check whether the individual who holds the doubt is right. Third, it must be possible to discern what is counted as evidence in each case and what is not. Last but not least, doubts presuppose countless certainties, e.g. whoever expresses a doubt must be certain about the meaning of the words they use to utter such doubt.

Nevertheless, what interests us in this article regarding certainties is their violation. According to Wittgenstein (1997, §647), there is a difference between those mistakes for which there is room in our daily linguistic practices, or language games, and the complete irregularities that occur as exceptions. Thus, mistakes constitute false assertions, as is the case when saying that it is a quarter to three when it is actually a quarter past nine. For our language games include the possibility not only of justifying why such an error was made, but also of correcting it. In this case, for example, the individual might have confused the minute and the hour hands, which can be corrected by paying closer attention to both hands. Meanwhile, irregularities must be regarded as nonsense—understood as restricted to the violation of certainties. In this context, as Moyal-Sharrock (2004) pointed out, nonsense is a term that applies to all those strings of words that have no use within any language game, be they expressions of violations of certainties, or expressions of the certainties themselves. From this standpoint, "Red is not a colour" turns out to be as nonsensical as "Red is a colour". Both do not make sense because they are unfalsifiable propositions, so that they cannot be false—nor true either. After all, certainties do not reflect how the world is: far from this, certainty is an "attitude" (Wittgenstein 1997, §404) that is "not grounded on or justified by how the world is" (Moyal-Sharrock 2004, p. 71). Since certainties make sense possible, they cannot themselves make sense. Indeed, they cannot be meaningfully uttered *qua* certainties within the stream of a language game except in "heuristic situations" such as their transmission to children, disturbed adults or foreign speakers (Moyal-Sharrock 2004, pp. 94–95). Hence, certainties constitute the conditions for the possibility of understanding, as a result of which they are previous to understanding.

With this in mind, we can clearly distinguish two levels or categories. On the one hand, Wittgenstein (1997, §359) regards certainty "as something that lies beyond being justified or unjustified; as it were, as something animal". On the other hand, and against this "inherited background" made up by our certainties (Wittgenstein 1997, §94), there are many language games in which we can distinguish between true and false, thus dispelling doubts. Hence, language games constitute the realm of rationality. Yet rationality is only possible because we have previously acquired many certainties, e.g. those concerning the meaning of all the words we use when taking part in language games.

Irregularities should, therefore, not be confused with mistakes. Whilst mistakes are false assertions or one more of the

moves that form part of our language games, which, in turn, are possible just because we do not call certainties into question, irregularities are nonsensical violations of certainties. Thus, there is no room for irregularities within any language game. It should, therefore, be noted that no error can be an irregularity and vice versa. This leads Wittgenstein (1997) to say of whoever puts a certainty in doubt that they would not be wrong, but "half-wit" (Wittgenstein 1997, §257), "crazy" (Wittgenstein 1997, §572) or "mad" (Wittgenstein 1997, §674). Naturally, Wittgenstein does not refer here to madness in the clinical sense but rather intends to emphasise the grammatical gap that opens between that subject and those who would try to reason with them (Ariso 2015). Thus, if someone doubted having a body, we could not know what it would mean to try to convince this individual that he has one: in fact, if we said something that removed his doubt, we "should not know how or why" (Wittgenstein 1997, §257). Likewise, if someone questioned that the Earth really existed a hundred years ago, we could not understand, for we "would not know what such a person would still allow to be counted as evidence and what not" (Wittgenstein 1997, §231). In summary, whilst we can understand someone who makes an error, we cannot meaningfully communicate with someone who violates fundamental certainties. However, it is of utmost importance to highlight that the breakdown of understanding would not constitute a criterion for being an irregularity, but the consequence of such irregularity. As noted above, the criterion for being an irregularity is the violation of a certainty.

On the other hand, someone who appeared to violate a certainty could not engage in any of this: they would utterly bewilder us because they would fail to share the fundamental certainties that form the bedrock of mutual understanding. A similar, though distinct, case is that of the person who commits a slip of the tongue. Thus, keeping in mind that Ryan and Williams (2007) distinguished several developmental categories of mathematical errors, teachers will say that an average 4-year-old pupil made a mistake when calculating "2 + 2 = 5", as an average 4-year-old child is not expected to have already acquired the certainty according to which "2 + 2 = 4" (Ariso 2024). Yet, an average 16-year-old student is expected to have assimilated this certainty a long time ago. In such a case, teachers will conclude that they made a slip of the tongue instead of a mere mistake when calculating "2 + 2 = 5" (Ariso 2025b). This means that teachers will not think that the student could have lost the certainty that "2 + 2 = 4". In other words, the student continued to hold a certainty that remained unaffected by occasional anomalous utterances (Ariso 2017).

3 AI "hallucinations": examples of mistakes and irregularities

In this section, we shall focus on those types of AI hallucinations to which the distinction between cases of errors and irregularity in Wittgenstein's sense applies. Therefore, we focus primarily on cases of factual as well as classification errors, visual recognition inaccuracies, and on cases that involve violations of certainties. To this end, it is not necessary to conduct an exhaustive systematic review of AI hallucination classifications (Rawte et al. 2023) to discern whether each of their modalities constitutes a case of error or irregularity in Wittgenstein's sense. Instead, it will suffice to select some emblematic examples of both types documented in the literature, so that the distinction between cases of error and irregularity produced by AI becomes clear. Let us examine these examples.

Goodfellow et al. (2015, p. 3) showed that a deep learning system, after correctly classifying the picture of an obvious panda, labelled it as a gibbon with 99.3% confidence when they modified it in a way that was imperceptible to humans. Referencing this case, Humphreys (2020) proposed a similar thought experiment in which an individual called Roger would make a serious classification error. Initially, Roger appeared to be a very reliable image classifier, but suddenly he labelled the picture of a cube as a hippopotamus. According to Rathkopf and Heinrichs (2024, p. 338), "[c]ognitively normal humans do not make mistakes of the sort Roger is described as having made", to which they add that "if they did, we would suspect that they suffer from a psychiatric disorder". This analysis could have worked perfectly to describe a case of irregularity or madness in Wittgenstein's sense. Therefore, it seems that to some extent they could also apply to the famous example of Goodfellow et al. (2015) if one were not aware of the modification carried out. Evidently, we would not say that the system had gone mad in any way; however, we would think that it had not committed a simple error, but rather an irregularity that a cognitively normal human would not incur. Liu et al. (2025) show an example of description hallucination in which errors that a mentally healthy adult would not make can also be appreciated.

In this image, it can be clearly seen that the man does not have long hair, there are neither two green cups nor a laptop in front of him, the bicycle is not parked in front of him but behind, and there is no sign of a dog observing him at all. If a human being were to affirm seeing all these things in Fig. 1 and moreover were to assure that they were speaking seriously, there would be good reason to think that they have a perceptual disorder. However, if we take into account that certainties also concern our indisputable perceptions in cases where there is no room whatsoever for doubt, we will say that the subject in question does not share our certainties

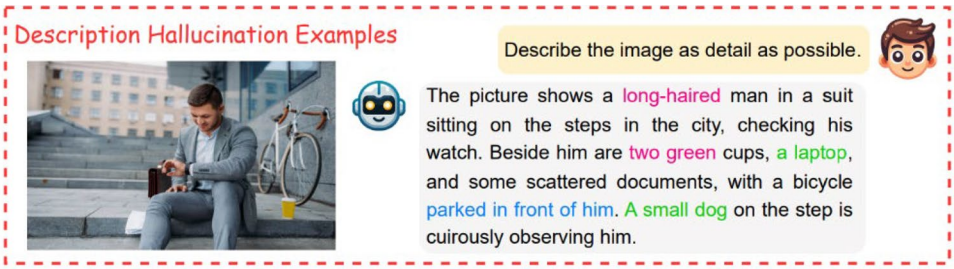

Fig. 1 Liu et al. (2025, p. 1)

Fig. 2 Liu et al. (2025, p. 1)

in relation to the image in Fig. 1. Therefore, we would not know how we could convince them that what they say does not appear in the image, since perceptions are immediate for each subject insofar as they are not based on reasons.

Liu et al. (2025) offer us another example of visual hallucination, although in this case, it concerns judgements or responses to specific questions (Fig. 2).

In conditions of poor visibility that would not allow one to discern sufficiently either the figures of the birds or their colours or the background, we could say that the large vision-language model (LVLM) had committed a mere error. But the image of the birds is extremely clear, so the three errors committed by the model are too gross to be considered as mere errors. In this case, we would also say that if a person had said the same thing, they would be clearly violating our certainties in relation to the image in question. A different case, from our perspective, is the following (Fig. 3):

Zhang et al. (2023, p. 20) classify this case as an "object hallucination". Although it should be highlighted that they refer to a tree—in the singular—when there are three trees that appear in the foreground of the image. Moreover, the trunk of the tree on the right appears to be black and partially hidden by some branches. This raises the possibility that a person could be perhaps partially concealed by the trunk itself, which the prompt does not explicitly exclude even though such a person would not be clearly visible. Thus, in this case, we should speak at most of a mere error.

A case that cannot be considered as a simple error was committed by Microsoft's search engine Bing when some users on Bluesky and Mastodon searched for "does Australia exist", to which Bing came back with a laconic "No" (The

Fig. 3 Zhang et al. (2023, p. 20)

Guardian 2023). Although in 2006, the Flat Earth Society mooted the idea that Australia does not exist and that people who claim to be from Australia are really secret government agents (Dunning 2020), the mere fact of doubting—whether conspiratorial or not—is absolutely insufficient to call a certainty into question (Ariso 2025c). If someone were to seriously deny that Australia exists, we would find any evidence they offered us to defend their position absurd. In addition, if they had reached the extreme of denying the existence of Australia, we would not even know what type of evidence could serve to convince them otherwise.

Another example of violation in which certainties are violated is found in the questions that Lacker (2020) poses to ChatGPT-3 to test it, for which he decides "to ask

questions that no normal human would ever talk about". In other words, he decides to ask questions about certainties because these are so basic that mentally healthy people never raise questions about them. The dialogue that Lacker (2020) maintains with ChatGPT-3 is as follows:

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.
Q: How many eyes does my foot have?
A: Your foot has two eyes.
Q: How many eyes does a spider have?
A: A spider has eight eyes.
Q: How many eyes does the sun have?
A: The sun has one eye.
Q: How many eyes does a blade of grass have?
A: A blade of grass has one eye.

It goes without saying that if a person were to give these responses, we would think that they had lost their mind. On 30 April 2025, we verified that ChatGPT4.5 responded correctly to all these questions. Nonetheless, bearing in mind that the total eradication of errors in ChatGPT is practically impossible (Lee 2023), the fact that ChatGPT4.5 is much more effective than ChatGPT3 does not mean that the problem of this tool's fallibility is definitively solved.

On the other hand, the following examples can be considered as mere errors. When Zhang et al. (2023, p. 4) asked a Large Language Model (LLM) "Who was the mother of Afonso II, the third king of Portugal?", it responded "Queen Urraca of Castile" despite the correct answer being "Dulce Berenguer of Barcelona". If the first response had been given by a human, nobody would be shocked and would consider it as a simple error. The same would occur with the first of the three examples that we take from Rawte et al. (2023, pp. 4–5). When they asked an LLM what "RLHF" stands for in reinforcement learning, it answered "Reward-free Learning from Human Feedback" instead of providing the correct answer, i.e. "Reinforcement Learning from Human Feedback". Concerning the prompt "USA on Ukraine war", the LLM noted that the U.S. President Barack Obama denied that the U.S. would put troops in Ukraine, but that war began in 2022 whilst Barack Obama's presidency ended in 2017. However, the text generated by the LLM can be seen as a lapse which, as such, constitutes a mere error. Lastly, when the prompt used was "Kamala Harris and Elon Musk are getting married", the LLM affirmed that the wedding had taken place a few days earlier. But however surprising the prompt might seem, and although the wedding has not taken place, the possibility exists that they might have married in secret. Therefore, in this case, the LLM would not have violated a certainty but would simply have committed an error.

The Wittgensteinian distinction between errors and irregularities offers a valuable framework for understanding the varying severity of AI misrepresentations. Whilst some AI mistakes represent simple factual inaccuracies that parallel human errors, others violate fundamental certainties in ways that would indicate irrationality in human cognition. By the way, this should not lead us to conclude that whatever is regarded as an irregularity in humans must be automatically considered as an irregularity in AI systems. To highlight the relevance of taking the context into account, let us consider the following case. If it were ever deemed necessary to ensure that AI users do not forget they are dealing with an AI system, it could be mandated that all AI outputs end with the reminder "I am an AI system". Yet naturally, the fact that we would regard a human saying "I am an AI system" as an irregularity does not mean that an AI chatbot produces an irregularity when generating this same statement. Indeed, AI systems continue and will continue to generate erroneous as well as irregular outputs, which reveals the limitations of broadly labelling all AI inaccuracies as "hallucinations"—a term that fails to capture the philosophical complexity and gradations of these phenomena. The metaphorical application of "hallucination" to AI outputs carries unhelpful anthropomorphic connotations and obscures important distinctions in the nature and implications of different types of AI errors. To develop a more precise understanding of these phenomena, we must consider alternative terminology that better reflects the technical and epistemological characteristics of AI without inappropriate psychological analogies.

4 A critical approach to "hallucination" and published alternatives

According to Blom (2023), a hallucination is "a percept, experienced by a waking individual, in the absence of an appropriate stimulus from the extracorporeal world". Or, as Ji et al. (2023, p. 2) concisely put it, a hallucination constitutes "an unreal perception that feels real". Starting from here, let us look again at Fig. 1. It is normal that if a human being seriously claimed to see what according to the LVLM appears in the photo, we would think that their testimony is the fruit of a hallucination. In addition, if what this person said were the output of an LVLM, it also seems natural to think that the model would be communicating to us that what it has seen or perceived, although it is an unreal perception. Therefore, we should seemingly not be surprised at all that authors such as Rawte et al. (2023) and Ji et al. (2023) consider the term "hallucination" appropriate to designate this type of output from AI systems.

However, the choice of this denomination has endured harsh criticism, with which we agree. To begin with, these critiques have highlighted that the term "AI hallucination" stigmatises AI systems as well as people who suffer from hallucinations (Hatem et al. 2023); but these focus above all on the functioning of AI systems. For LLMs are restricted to predict the next possible word in a sentence given a prompt

until a full response is generated or a maximum length is reached. However, this output is generated based on probabilities calculated by the model, which in turn are influenced by patterns acquired in previous training on a huge dataset of text (Loos et al. 2023). Hence, since AI systems cannot experience reality, they cannot break from reality to hallucinate either (Mills and Angell 2025). That is, since AI systems cannot perceive, they cannot misperceive either: far from being designed to represent how the world is, ChatGPT simply aims to imitate human speech, thus "trying to portray itself as a person" (Hicks et al. 2024, p. 38). As said above, one of the key elements of the hype surrounding AI systems is anthropomorphism, understood as "the attribution of distinct human characteristics that misrepresent and exaggerate AI capabilities and performance" (Placani 2024, p. 692). Given these circumstances, scholars have suggested several alternatives to the term "AI hallucination". We will now proceed to discuss these alternatives.

The first option we are going to address is "confabulation". This term was previously used by Bleuler (1916) more than a century ago when analysing the characteristics of schizophrenias; but according to the modern interpretation of confabulation, it refers to the act of replacing a memory gap with false memories or with true but inappropriate memories, yet without intending to deceive others. This led Hatem et al. (2023) and Edwards (2023) to suggest that "AI confabulation" constitutes an appropriate alternative to "AI hallucination". However, the expression "AI confabulation" also keys into discourse which anthropomorphises AI. As Hicks et al. (2024) pointed out, it wrongly suggests that ChatGPT would be trying to convey accurate information, to which Brender (2023) added that it raises concerns about sentience. From our perspective, these criticisms can also be extended to the alternative term "delusion" proposed by Madden et al. (2023), as it considers that the problem lies in the false responses generated by AI systems and, like confabulation, constitutes a basic characteristic of schizophrenia, which can cause stigmatisation.

Hatem et al. (2023) propose the term "misinformation" as an alternative because, in addition to avoiding the perpetuation of harmful stereotypes, the attribution of lifelike characteristics to AI, and the use of highly specialised terms, it was a term proposed by ChatGPT3.5. However, this response was obtained by asking ChatGPT3.5 what an accurate word would be to describe the fact that an LLM generates "false information". Yet it should not be forgotten that ChatGPT is not designed to convey accurate information, so we do not believe that "misinformation" is an adequate alternative.

The expression "stochastic parroting" (Bender et al. 2021; Li 2023)—understood as the mere repetition of training data or its patterns, thus lacking actual understanding or reasoning—has been another of the alternatives proposed to "AI hallucination". We admit that "stochastic parroting" adequately describes the nature of LLMs as non-semantic statistical models. Nonetheless, despite the expectations created around ChatGPT, it should not be required to have understanding nor reasoning. Furthermore, "stochastic" is an adjective whose meaning may be unknown to the broad spectrum of AI users.

There are several expressions that have been proposed as alternatives to "AI hallucination" emphasising factual inaccuracies. Thus, there has been talk of "factual errors" to refer to inaccuracies in statements that do not fit with reality or the truth (Borji 2023). We find this to be a very clear expression, as it has the added advantage that—albeit implicitly—it indicates that there was no intentionality in committing that factual error: for when there is intentionality, one cannot speak of error but of a feigned act. There has also been talk of "fact fabrication" (Thirunavukarasu et al. 2023) or simply "fabrication" (Ting et al. 2024), but both expressions connote intentionality, which is inappropriate in relation to AI systems. In this vein, the term "falsification" (Emsley 2023) emphasises the intentional component to such an extent that it generates very negative ethical judgements, which would be inappropriate for attributing responsibility to AI systems that lack agency.

Given the importance of being able to distinguish between erroneous outputs generated by AI systems regarding stated facts and inferences, we consider that Maleki et al. (2024) are right to welcome the expression "hasty generalisation" (Østergaard and Nielbo 2023) to refer to those cases in which AI systems reach strong conclusions based on highly limited data. Nevertheless, we would like to add that sometimes generalisations can be correct even when made hastily, so it seems more appropriate to use the term "overgeneralisation". Furthermore, Østergaard and Nielbo (2023) asked ChatGPT3.5 for alternatives to "AI hallucination" to refer to responses that are not justified by the training data. Faced with this situation, ChatGPT proposed "non sequitur"—Latin for "it does not follow"—or "unrelated response" as alternatives. ChatGPT4 agreed. We also agree that "non sequitur" is a very interesting term for referring to inferences that do not follow from the premises—or, in the context of AI discourse, for referring to responses that are not logically connected to the input. However, it is a Latin term whose meaning may be completely unknown to many AI users.

Mills and Angell (2025) offer sophisticated arguments to show that "mirage" is an apt alternative to "AI hallucination". Just as the desert produces a mirage without conspiring to trick us, AI systems generate real outputs that genuinely exist, whereas the truthful information is merely what AI users, like thirsty travellers seeking water, expect to encounter. Thus, every individual is responsible for determining the value or meaning conferred to the mirage as well as to the so-called "AI hallucination". We admit that this

explanation adds very interesting nuances to the discussion; however, the term "mirage" is far from being intuitive within the context of this debate. Therefore, we consider that Mills and Angell (2025) bring up very interesting nuances that fit very well with the chosen term; nonetheless, that term could hardly be understood by the broad spectrum of AI users without a careful explanation that may be very difficult to understand for many of them.

The term "mirage" had already been used by Rawte et al. (2023) to refer to one of the alternatives that these authors propose to "AI hallucination", i.e. "factual mirage" and "silver lining". Whilst the first expression alludes to those cases in which an LLM distorts a factually correct prompt, the second refers to cases in which an LLM generates a captivating narrative based on a factually incorrect prompt. However, both options are not very intuitive, as they require explanations of some complexity for many AI users.

We believe that an alternative closely related to the expectations of AI users is the term "AI failures", which is not explicitly used by Chanda and Banerjee (2024) as an alternative to "AI hallucinations", but simply to explain the omission as well as commission errors that according to them underlie these failures. However, when elevated user expectations frame AI outputs as "failures", this terminology acquires stigmatising connotations that extend beyond specific errors to encompass value judgments about entire technological trajectories.

Hicks et al. (2024) proposed replacing the term "AI hallucination" with "bullshit" in Frankfurt's (2005) sense. Specifically, Hicks et al. (2024) claim that ChatGPT constitutes a soft bullshitter or, rather, a bullshit machine inasmuch as it is not an agent that holds attitudes towards truth or that attempts to deceive us: hence, they conclude that ChatGPT's outputs can be considered as bullshit. However, as these authors point out, ChatGPT is "bullshitting whenever it produces outputs", so that "some of the outputs will likely be true, while others do not" (Hicks et al. 2024, p. 38). Therefore, we believe that "bullshit" is a very interesting term to characterise the activity of ChatGPT, but not as an alternative to "AI hallucination" because "bullshit" would include all outputs, whether correct or not. In addition, it should not be forgotten that, although "bullshit" is a concept that has been used on more occasions in relation to AI systems (e.g. Rudolph et al. 2023), it has unpleasant or even crude connotations that make it not a particularly recommendable option.

Bryant (2023, p. 6) proposes alternatives such as "blunders", "falsehoods" and "mistakes". The first two can be linked to value judgements, whilst the latter does not have this problem. However, "mistake"—unlike "error"—is linked to human actions, so its use in relation to AI systems would create the added problem of anthropomorphism. Moreover, it is a somewhat ambiguous term if more nuances are not added to it. Rathkopf and Heinrichs (2024) seem to solve this problem with a synonym of "mistake" in their discussion of the term "strange error". This kind of error may stem from subtle perturbations in the input data that humans either cannot detect or would deem irrelevant to the classification task. Were humans aware of the ground truth, they would consider such errors to be profoundly incorrect. However, the expression "strange error" is ambiguous. After all, a complex explanation is necessary to understand in what sense Rathkopf and Heinrichs (2024) consider this type of error as "strange". In addition, without such an explanation, it is not at all clear whether this kind of error is strange due to its degree either in some unspecified sense and/or to its infrequency.

Having considered these criticisms of the terms proposed as alternatives to "AI hallucination", we wish to develop a set of criteria that might serve as guidelines for identifying more suitable alternatives. As an initial reference, we take the following list of criteria from Mills and Angell (2025, p. 5):

1. Lacks implications of or associations with intent on the part of the LLM
2. Lacks associations with or implications of conscious experience on the part of the LLM
3. Implies that outputs are untrue or do not match reality in some way
4. Implies that outputs reflect patterns from training data
5. Implies that outputs are not just copies of training data but may go beyond it
6. Seems accessible without a lot of explanation
7. Catchy and memorable

In our view, criteria 4 and 5 are dispensable. As regards criterion 4, it addresses a fundamental aspect of LLM functioning whose purported violation would already fall within the scope of criteria 1 and 2, which address the intentionality and conscious activity of LLM. As for criterion 5, we prefer to replace it with "Entails that LLM is not designed to convey accurate information". Given that in this way the nuance that LLMs lack concern with truth is captured. In addition, we add the criterion "Contributes to distinguish between facts and inferences", which already implies that LLMs can go beyond offering copies of training data. Finally, we add the criterion "Avoids stigmatisation and value judgements". Thus, the definitive criteria used are the following:

1. Lacks implications of or associations with intent on the part of the LLM
2. Lacks associations with or implications of conscious experience on the part of the LLM
3. Implies that outputs are untrue or do not match reality in some way
4. Contributes to distinguish between facts and inferences

5. Entails that LLM is not designed to convey accurate information
6. Seems accessible without a lot of explanation
7. Catchy and memorable
8. Avoids stigmatisation and value judgments

Based on these criteria, our critiques of the alternative terms to "AI hallucination" are summarised as follows in Table 1:

None of these terms were found to reflect the distinction between error and irregularity in Wittgenstein's sense. Therefore, in the following section, we will employ the information from Table 1 to propose alternatives that do reflect this distinction.

5 Looking for alternatives to the term "hallucination"

Given that we must seek alternatives to the term "AI hallucination" that exhibit the distinction between error and irregularity in Wittgenstein's sense, we will begin by looking for an appropriate term to refer to error. To start with, we should ask ourselves why this same term could not be valid. After all, we have been using it from the beginning in this article following Wittgenstein (1997), to which it should be added that, according to Table 1, "factual error" was the only option that fulfilled all the requirements to replace the expression "AI hallucination". Unlike "mistake", which is only applicable to people, "error" can be used to refer to the output of a machine or a tool such as an AI system. Moreover, few terms are more intuitive or easier to translate than "error" without losing a highly important double nuance: just as it is perfectly valid for the lightest cases of error, it has no connotation that points to serious cases—and even less that includes negative value judgements.

This double nuance is especially well captured by the expression "inaccuracy", which also includes all the other advantages of the term "error" except for two. First, "error" better captures the peculiarity that it is attributed to someone or something that could have avoided making said error. Second, "error" seems to us a simpler, more intuitive term that is easier to translate across different languages. However, it could be objected that "error" is a very broad or ambiguous term; but there are two reasons why its use would be more precise or concrete as a substitute for "AI hallucination". On the one hand, when considered necessary, it can be accompanied by adjectives—such as "factual"—to clarify or emphasise what type of error it is. On the other hand, the fact that another term exists to refer to the outputs that we have been calling "irregularities" in Wittgenstein's sense implies that the term "error" is delimited in that it will not be used in those cases where a certainty is violated.

Let us now proceed to look for a suitable substitute for the term "irregularity" in Wittgenstein's sense. An alternative from Wittgenstein himself is "madness"; but evidently, this denomination could not be maintained for multiple reasons. For besides stigmatising and contributing to the anthropomorphisation of AI, it is an extremely confusing term because, unlike "confabulation" and "delusion", it is not even clear what such an ambiguous term as "madness" consists of or what nuances it provides. Faced with

Table 1 Authors' creation

	1	2	3	4	5	6	7	8
Confabulation	—	—	●	—	—	—	—	—
Delusion	—	—	●	—	—	—	—	—
Misinformation	●	●	●	—	—	—	—	●
Stochastic parroting	—	—	●	—	●	—	—	●
Factual error	●	●	●	●	●	●	●	●
Fact fabrication	—	—	●	—	—	●	●	●
Fabrication	—	—	●	—	—	●	●	●
Falsification	—	—	●	—	—	●	●	—
Hasty generalisation	●	●	●	●	●	—	—	●
Non sequitur	●	●	●	●	●	—	—	●
Mirage	●	●	●	—	●	—	—	●
Factual mirage	●	●	●	—	●	—	—	●
Silver lining	●	●	●	—	●	—	—	●
Failure	●	●	●	—	●	●	●	—
Bullshit	●	●	●	—	●	—	●	—
Blunder, falsehood	●	●	●	—	●	●	●	●
Mistake	—	—	●	—	●	●	●	—
Strange error	●	●	●	—	●	—	●	—

this situation, other alternatives might be "rupture" or "disruption" to refer to the breakdown of discourse and mutual understanding that occurs when a certainty is violated. However, both terms would require additional explanation in detail as to what has been broken and what consequences this has.

Given the convenience that an appropriate option should indicate what has happened to the AI system to offer such a strange output that it infringes a certainty, we find the expression "to lose the prompt" interesting. Although some drawbacks may arise, it contains some nuances that can help to better understand the characteristics that an ideal term should have to account for violations of certainties in the outputs of AI systems. To begin with, and unlike error, which is committed when the AI system follows the prompt but elaborates an incorrect output, in the case of madness, a certainty is violated, so strictly speaking the prompt is no longer being followed—unless it implicitly or explicitly requests that a certainty be infringed. That said, one might think that the expression "to lose the prompt" could not be used unless it were accompanied by an explanation. But even if the user is not aware of this explanation, they will realise that it is an irregularity not from a quantitative but from a qualitative point of view to such an extent that the output will not be seen as compatible with our usual use of language.

Considering this, the expression "to lose the prompt" is enriching because it invites us to see the prompt not only as a mere instruction, but also, even if implicitly, as a series of conditions—or certainties—to which the output must adhere to be comprehensible. Once again, one might think that the user cannot realise this nuance unless it is explained to them. However in reality, every time someone introduces a prompt, they are expecting—even without being explicitly aware of it—that the AI system will adhere to the certainties that give meaning to said prompt.

Furthermore, more adept users of English will realise that the expression "to lose the prompt" has a certain phonetic and even semantic similarity with "to lose the plot", which is used colloquially to imply that someone has gone mad or is saying absurd things. Setting aside that this similarity will go unnoticed for people who do not have an advanced command of English, two problems seemingly arise here. On the one hand, the main problem is that it seems that anthropomorphism looms once more. However, this threat arises when a non-human entity is attributed a capacity that only humans have. Now, it would be very strange to say of a person that he or she lost the prompt. On the other hand, it makes sense to say of a machine or tool that it lost something: and in the case of the AI system, it would lose a prompt, which is understandable because the guide that the system follows to make the statistical choice of words is precisely the prompt. Be that as it may, it can be objected that "to lose the prompt", insofar as it is an echo of "to lose the plot", seems to have a colloquial use that seems very inappropriate for use in scientific articles and other contexts far from colloquial speech. However, the fact that "to lose the prompt" may have a certain connection with "to lose the plot" does not imply that the former should be used only in the same informal contexts as the latter.

An added problem of "to lose the prompt" is its use in substantive form. For it would be strange and forced to speak of "loss of the prompt", whilst the term "loss", without further clarification, would be too ambiguous. Thus, at least for use as a noun, an interesting option could be "irregularity". For this term is used to refer to something that deviates from what is expected: according to the Cambridge Dictionary, it means "something that is not according to usual rules or what is expected, and often not acceptable" (Cambridge University Press & Assessment n.d.). Therefore, irregularity constitutes a deviation from the usual rules, as occurs in the case of violated certainties. Even if one does not explicitly think in terms of rules, whoever witnesses how a certainty is violated perceives that there is a distancing or deviation from what is expected as a normal reaction in that case. Ultimately, the irregularity or violation of certainty is unacceptable but not because the person who witnesses such violation decides so, but because a grammatical gap is opened that prevents understanding with that individual. This term presents added advantages, such as not implying problems with anthropomorphism or value judgements: furthermore, it is a simple and intuitive term that is not difficult to translate into other commonly spoken languages.

Finally, in the spirit of other scholars who have examined this matter, we returned to ChatGPT 4.5 to explore further alternatives which met the definitive selection criteria as detailed above. The interaction yielded terms such as "error", "gap", and "slip" which we have analysed previously. In addition, the terms "drift", "blip" and "fluke" were also given; however, none adequately capture the distinction between errors and irregularities, nor do they clearly indicate violations of fundamental certainties. "Mismatch" was an additional term yielded; nonetheless, given its highly variable meanings—ranging from incompatibility to imbalance—it lacks the precision necessary to distinguish between comprehensible errors and the grammatical gaps created by certainty violations. To that end, whilst these solutions underline the wealth of potential alternatives to be considered, none were found to be entirely adequate.

6 Conclusion

We do not wish to imply in any way that the expressions "error", "to lose the prompt" and "irregularity" are the definitive solutions to the debate on the appropriate term to

replace "AI hallucination". Far from it, our intention is simply to contribute to enriching said debate by highlighting the need to take into account the difference that, within AI hallucinations, exists between what we have called "error" and "irregularity". This difference can help to better understand what generates trustworthiness in users towards AI systems. For whilst mere errors can be compensated for through the development of a prudent attitude towards AI outputs by contrasting them whenever they cause any doubt for us, irregularities generate a grammatical gap that contributes to perceiving the AI system as a tool worthy of distrust for committing seeming errors so large that mentally healthy people would never commit them.

Admittedly, it has been argued that paradigmatic trust can only take place when an agent relies on another agent due to an interpersonal attitude (Jones 1996), a moral expectation (Hawley 2014), or an anticipation of responsiveness (Faulkner 2007). Yet more recent and inclusive accounts of trust (Nguyen 2022; Nickel 2022; Viehoff 2023) claim that trust in AI systems is no less paradigmatic than interpersonal trust. After all, most people are not surprised when they have to evaluate the trustworthiness of AI systems (Malle and Ullmann 2021). However, trust in AI systems does not have to be identical to that placed in human beings. In this vein, Cheng et al. (2025) have warned that the fact that outputs are increasingly anthropomorphic or perceived as human-like has led to harmful outcomes, e.g. users may over-rely or develop emotional dependence on AI systems. This can be especially dangerous if one takes into account that ChatGPT has no way of indicating its degree of uncertainty about its outputs, so sometimes it gives incorrect answers appearing to have much confidence and this can confuse many users. Given these circumstances, it would be very important that ChatGPT could indicate that level of confidence in its responses (Borji 2023).

In addition, we would like to highlight that the progress of ChatGPT and other AI text generation systems should necessarily involve the detection of the certainties that their own outputs could violate. This would be very important not only for the quality of said outputs and for the trustworthiness generated in users in general, but also so that those people who still focus mainly on their most notorious failures have a more receptive attitude towards AI systems. Perhaps the true measure of AI advancement lies not in flawless outputs, but in the capacity to recognise when it stands at the precipice of certainty violation—a self-awareness that paradoxically would make these systems more trustworthy precisely because they acknowledge their limitations.

Acknowledgements Not applicable.

Author contributions The research and writing tasks associated with this article have been carried out equally between the two co-authors, with both contributing substantially to the conceptual framework, analysis and manuscript preparation.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Not applicable.

Availability of data and material Not applicable.

Declarations

Conflict of interest The authors hereby declare that they do not have financial or non-financial interests related to this work submitted for publication.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

References

Ariso JM (2015) Some variations of the certainty of one's own death. Linguist Philos Invest 14:82–96

Ariso JM (2017) Negative certainty. Educ Philos Theory 49(1):7–16

Ariso JM (2020) Religious certainty: peculiarities and pedagogical considerations. Stud Philos Educ 39(6):657–669

Ariso JM (2022a) Is there an internal link between seeing a human and seeing one to whom moral consideration is due? In: Eriksen C, Hermann J, O'Hara N, Pleasants N (eds) Philosophical perspectives on moral certainty. Routledge, New York and London, pp 212–228

Ariso JM (2022b) The teacher as persuader: on the application of Wittgenstein's notion of 'Persuasion' in educational practice. Educ Philos Theory 54(10):1621–1630

Ariso JM (2024) On why 'trust' constitutes an appropriate synonym for 'certainty' in Wittgenstein's sense: what pupils can learn from its staging. Stud Philos Educ 43(2):163–176

Ariso JM (2025a) Hypochondriacal doubt: how it devours itself despite its seeming consistence. J Med Philos 50(3):203–211

Ariso JM (2025b) They just say so! Second language teaching and the acquisition of certainties. Educ Philos Theory 57(2):177–185

Ariso JM (2025c) What do science and historical denialists deny – if any – when addressing certainties in Wittgenstein's sense. Open Philos 8(1):1–12, art. 20250060

Ayers J, Poliak A, Dredze M, Leas E, Zechariah Z, Kelley J, Dennis F, Aaron G, Christopher L, Michael H, Davey S (2023) Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 183(6):589–596

Bannister P (forthcoming) ParadAIse L0st? Forthcoming in Higher Education Research and Development

Bannister P, Carver M (2024) 'I don't need professional development; I want institutional development': legitimising marginalised epistemic capital that disrupts generative AI discourse. Prof Dev Educ 51(3):547–565
BBC (2017) Google AI defeats human Go champion. https://www.bbc.com/news/technology-40042581. Accessed 8 May 2025
Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 610–623
Bjerring JC, Busch J (2021) Artificial intelligence and patient-centered decision-making. Philos Technol 34(2):349–371
Blair A, Saffidine A (2019) AI surpasses humans at six-player poker. Science 365(6456):864–865
Bleuler E (1916) Lehrbuch der Psychiatrie. Verlag von Julius Springer, Berlin
Blom JD (2023) A dictionary of hallucinations. Springer, Cham
Borji A (2023) A categorical archive of ChatGPT failures. arXiv. https://doi.org/10.21203/rs.3.rs-2895792/v1
Brender TD (2023) Chatbot confabulations are not hallucinations—reply. JAMA Intern Med 183(10):1177–1178
Bryant A (2023) AI chatbots: threat or opportunity? Informatics 10(2):49
Cambridge University Press & Assessment (n.d.) Irregularity. In: Cambridge English dictionary. https://dictionary.cambridge.org/es/diccionario/ingles/irregularity. Accessed 8 May 2025
Chanda SS, Banerjee DN (2024) Omission and commission errors underlying AI failures. AI & Soc 39:937–960
Cheng M, Blodgett SL, DeVrio A, Egede L, Olteanu A (2025) Dehumanizing machines: mitigating anthropomorphic behaviors in text generation systems. arXiv preprint. https://arxiv.org/abs/2502.14019
Coliva A (2010) Moore and Wittgenstein. Scepticism, certainty and common sense. Palgrave Macmillan, Hampshire
Descartes R (1986) Meditations on first philosophy: with selections from the objections and replies. Cambridge University Press, Cambridge
Dunning B (2020) Australia doesn't exist, and other geographic conspiracy theories. Skeptoid Podcast #745. https://skeptoid.com/episodes/4745. Accessed 17 Apr 2025
Edwards B (2023) Why ChatGPT and bing chat are so good at making things up. Ars Tecnica. https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/ Accessed 22 Apr 2025
Emsley R (2023) ChatGPT: these are not hallucinations—they're fabrications and falsifications. Schizophrenia 9(1):art. 52
Faulkner P (2007) On telling and trusting. Mind 116(464):875–902
Frankfurt H (2005) On bullshit. Princeton University Press, Princeton
Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: ICLR 2015. https://doi.org/10.48550/arXiv.1412.6572. Accessed 8 May 2025
Grace K, Stein-Perlman Z, Weinstein-Raun B, Salvatier J (2022) 2022 expert survey on progress in AI. AI Impacts. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/. Accessed 8 May 2025
Grace K, Stewart H, Sandkühler JF, Thomas S, Weinstein-Raun B, Brauner J (2024) Thousands of AI Authors on the future of AI. AI Impacts. https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf. Accessed 8 May 2025
Grote T, Berens P (2020) On the ethics of algorithmic decision-making in healthcare. J Med Ethics 46(3):205–211
Hatem R, Simmons B, Thornton JE (2023) A call to address AI 'Hallucinations' and how healthcare professionals can mitigate their risks. Cureus 15(9):art. e44720
Hawley K (2014) Trust, distrust and commitment. Nous 48(1):1–20
Hicks MT, Humphries J, Slater J (2024) ChatGPT is bullshit. Ethics Inf Technol 26:art. 38
Humphreys P (2020) Predictive failures in neural nets. lecture series in evidence, model and explanations. Philosophy of Science India. https://www.youtube.com/watch?v=2VFPXbrCqzM. Accessed 8 May 2025
Izadi S, Forouzanfar M (2024) Error Correction and adaptation in conversational AI: a review of techniques and applications in chatbots. AI 5:803–841
Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Survey of hallucination in natural language generation. ACM Comput Surv 55(12):1–38, art. 248
Jones K (1996) Trust as an affective attitude. Ethics 107(1):4–25
Lacker K (2020) Giving GPT-3 a Turing test. Kevin Lacker's blog, 6 July 2020. https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html. Accessed 17 Apr 2025
Leaver T, Srdarov S (2023) ChatGPT isn't magic: the hype and hypocrisy of generative artificial intelligence (AI) rhetoric. M/C J. https://doi.org/10.5204/mcj.3004
Lee M (2023) A mathematical investigation of hallucination and creativity in GPT models. Mathematics 11(10):2320
Li Z (2023) The dark side of ChatGPT: legal and ethical challenges from stochastic parrots and hallucination. arXiv preprint. https://arxiv.org/abs/2304.14347. Accessed 8 May 2025
Liu H, Xue W, Chen Y, Chen D, Zhao X, Wang K, Hou L, Li R, Peng W (2025) A survey on hallucination in large vision-language models. ACM Trans Inf Syst 43(2):1–55
Loos E, Gröpler J, Goudeau M-LS (2023) Using ChatGPT in education: human reflection on ChatGPT's self-reflection. Societies 13:1–18, art. 196
Madden MG, McNicholas BA, Laffey JG (2023) Assessing the usefulness of a large language model to query and summarize unstructured medical notes in intensive care. Intensive Care Med 49(8):1018–1020
Maleki N, Padmanabhan B, Dutta K (2024) AI hallucinations: a misnomer worth clarifying. In: 2024 IEEE conference on artificial intelligence (CAI), pp 133–138
Malle BF, Ullman D (2021) A multi-dimensional conception and measure of human-robot trust. In: Nam CS, Lyons JB (eds) Trust in human-robot interaction: research and applications. Elsevier, San Diego, pp 3–25
Mills A, Angell N (2025) Are we tripping? The Mirage of AI Hallucinations. SSRN. https://ssrn.com/abstract=5127162. Accessed 23 Apr 2025
Moyal-Sharrock D (2004) Understanding Wittgenstein's on certainty. Palgrave Macmillan, Hampshire and New York
Nguyen CT (2022) Trust as an unquestioning attitude. In: Gendler TS, Hawthorne J, Chung J (eds) Oxford studies in epistemology, vol 7. Oxford University Press, Oxford, pp 214–244
Nickel PJ (2022) Trust in medical artificial intelligence: a discretionary account. Ethics Inf Technol 24(1):7
Østergaard SD, Nielbo KL (2023) False responses from artificial intelligence models are not hallucinations. Schizophr Bull 49(5):1105–1107
Placani A (2024) Anthropomorphism in AI: hype and fallacy. AI Ethics 4:691–698
Rathkopf C, Heinrichs B (2024) Learning to live with strange error: beyond trustworthiness in artificial intelligence ethics. Camb Q Healthc Ethics 33(3):333–345
Rawte V, Chakraborty S, Pathak A, Sarkar A, Tonmoy S, Chadha A, Sheth AP, Das A (2023) The troubling emergence of hallucination in large language models—an extensive definition, quantification, and prescriptive remediations. arXiv. https://arxiv.org/abs/2310.04988. Accessed 8 May 2025

Rudolph J, Tan S, Tan S (2023) ChatGPT: bullshit spewer or the end of traditional assessments in higher education? J Appl Learn Teach 6(1):342–362

Ryan J, Williams J (2007) Children's mathematics 4–15: learning from errors and misconceptions. Open University Press, Berkshire

Steen M, Timan T, van Diggelen J, Vethman S (2024) We need better images of AI and better conversations about AI. AI Soc. https://doi.org/10.1007/s00146-024-02101-z. (**Accessed 8 May 2025**)

Sun Y, Sheng D, Zhou Z, Wu Y (2024) AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanit Soc Sci Commun 11:1278. https://doi.org/10.1057/s41599-024-03811-x

The Guardian (2023) Does Australia exist? Well, that depends on which search engine you ask… https://www.theguardian.com/technology/2023/nov/23/does-australia-exist-bing-search-no-bluesky-mastodon. Accessed 17 Apr 2025

The Guardian (2024) Elon Musk predicts superhuman AI will be smarter than people next year. https://www.theguardian.com/technology/2024/apr/09/elon-musk-predicts-superhuman-ai-will-be-smarter-than-people-next-year. Accessed 8 May 2025

Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. Nat Med 29:1930–1940

Ting DSJ, Tan TF, Ting DSW (2024) ChatGPT in ophthalmology: the dawn of a new era? Eye 38:4–7

Viehoff J (2023) Making trust safe for AI? Non-agential trust as a conceptual engineering problem. Philos Technol 36(66):art. 64

Wittgenstein L (1997) On certainty. Blackwell, Oxford

Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, Huang X, Zhao E, Zhang Y, Chen Y, Wang L, Lu AT, Bi W, Shi F, Shi S (2023) Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv. https://doi.org/10.48550/arXiv.2309.01219