Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

What is AI Hallucination? Defining the Term "Hallucination" in the Context of AI Literature and Discourse through a Literary Review of Existing Works

Kristopher Michael Sandoval

[1]Research and Development, DeveloperFirst

Correspondence
Corresponding author Kristopher Sandoval.
Email: kristopherleads@gmail.com

Abstract

In discourse surrounding Artificial Intelligence (AI) and Large Language Models (LLMs), a phenomenon has been observed which has been termed 'hallucination'. Generally speaking, this term has come to represent a general 'catch-all' for any time a model generates an output which is deemed 'inaccurate'. Despite its broad use in both academic and general conversation, the use of the term 'hallucination' is problematic due to its overly vague definition - the term is often used to represent very different kinds of inaccuracies. In this work, we review the existing literature in the field of Artificial Intelligence and Large Language Models to ascertain how the term 'hallucination' is used, and what co-associated terminology can be extracted for better contextualization. Based upon our findings, we suggest five core types of 'hallucination' that can be referred to directly as a means to disambiguate the type of 'hallucination' which is being observed in any given state. These five types are Hallucination by Lacking Context, Hallucination by Stochasticism, Hallucination by Epistemology, Hallucination by Survivalism, and Hallucination by Intention.

KEYWORDS
Large Language Models, Artificial Intelligence, Hallucination

1 | INTRODUCTION

The use of the term 'hallucination' in artificial intelligence is largely based on previous work focused on the spontaneous generation of new ideas in computing systems which are not associated with contextual or metacontextual content.

Tait (1982) observed spontaneous errant generation of text summaries which did not represent the core idea of the body text while developing a computer program called Scrabble . While this was an early example of hallucinatory behavior in computer systems, it was not directly referred to as 'hallucination' at that time. In 1985, a doctoral dissertation by Eric Daniel Mjolsness directly used the term 'hallucination' to refer to the generation of simulated fingerprints from random input data. Mjolsness (1986) is thus one of the earliest examples of the term 'hallucination' being used in the context which would define it in modern discourse within Artificial Intelligence and Large Language Models.

Later work, such as Thaler (1995), theorized that this spontaneous generation arose from internal pattern recognition systems attempting to resolve unrelated input into a previously observed pattern. Thaler referred to this occurrence within virtual reality systems as 'virtual input'. Another work, Baker and Kanade (2000), observed additional hallucinatory behavior in complex models. According to their report on an algorithm to improve the resolution of an image for use in facial recognition and license plate reading, they noted that the extra pixels generated for resolution improvements were effectively 'hallucinated', as they derived from suggested patterns rather than direct input.

The terminology 'hallucination' arguably entered the broader technological lexicon in relation to artificial intelligence with the release of a Google Research paper titled "Hallucinations in Neural Machine Translation". Google (2018) was backed and disseminated by Google, a leading technology provider and significant investor in the Artificial Intelligence space; as such, it is an influential use case of the term in modern academic discourse.

Over time, the terminology 'hallucination' has evolved broadly in terms of both definition and applicability in the academic and technological space. The term has come to be a general 'catch-all' for any output generated by an automated system which seems to be inaccurate, out-of-context, or otherwise unpredicted. Due to this broad and

Abbreviations: AI, artificial intelligence; LLM, large language models.

vague definition, 'hallucination' has come to mean many different things, lacking structured continuity or formal definition.

A survey of existing literature shows that 'hallucination' has come to represent vastly different things.

In Monteith et al. (2024), an article in The British Journal of Psychiatry, hallucination is referred to as a word used to "describe output generated by LLM that is nonsensical, not factual, unfaithful to the underlying content, misleading, or partially or totally incorrect". This is a generalized definition, and does nothing to indicate why the 'hallucination' is occurring, the nature of the inaccuracy, or the severity of that inaccuracy.

In another article in the Curues journal, Athaluri et al. (2023), hallucination was broadly defined as "a phenomenon where AI generates a convincing but completely made-up answer". This definition narrowly defines 'hallucination' to be content which is convincing yet entirely made-up, a state which is not always the case.

In Rawte et al. (2023), a study with the AI Institute at the University of South Carolina, a broader definition suggests that hallucination is "the generation of content that strays from factual reality or includes fabricated in- formation". This definition misses hallucinatory behavior which can generate answers which may be correct in the context of the original data set but which is irrelevant or otherwise non-contextual to the question being asked.

Generally speaking, many definitions of hallucination in the context of Artificial Intelligence fall into one of two categories:

- The definition is too narrowly defined, serving only to provide a basic term to discuss a specific example or kind of hallucination that is being covered in the specific discourse at hand; or
- The definition is to broadly defined, allowing for broad-stroke statements and generalizations that are somewhat accurate in their totality but miss nuances and important caveats for each given kind of hallucination.

This issue of nomenclature has caused significant deviation in the meaning of hallucination in this context, introducing confusion in discourse, additional contextual provisions, clarifications, and other issues whenever the term is used.

To help mitigate this issue, we have conducted a brief survey of the common use of the term 'hallucination' in academic literature on this topic. From this survey data, we have created a suggested framework by which hallucination can be understood and re-framed to give additional context in the immediate use.

2 | METHODOLOGY

In order to offer some potential specific sub-definitions for the term 'hallucination', we must first view the incidence of hallucination and related terms across the literature. To do this, we surveyed several academic databases for content focused on artificial intelligence hallucination. The multidisciplinary databases queried were:

- Google Scholar (accessed via https://scholar.google.com/)
- EBSCO Open Dissertations (accessed via https://opendissertations.org/)
- JSTOR (accessed via https://www.jstor.org/)
- PubMed (accessed via https://pubmed.ncbi.nlm.nih.gov/)
- ERIC (accessed via https://eric.ed.gov/)
- IEEE Xplore (accessed via https://ieeexplore.ieee.org/Xplore/home.jsp)
- Science Direct (accessed via https://www.sciencedirect.com/)
- DOAJ (accessed via https://doaj.org/, Articles feature)

From this list of databases, we queried for content related to hallucination using the string '"artificial intelligence" AND "hallucination"'. This string allowed us to extract content where both 'artificial intellignece' and 'hallucination' are both used, representing the broader corpus of academic content on this topic.

The values from this string search were then compiled, resulting in the data shown in Table I.

TABLE 1 Incidence of "Artificial Intelligence" AND "Hallucination" in Mainstream Academic Databases

Google Scholar	25,500
EBSCO Open Dissertations	3
JSTOR	432
PubMed	61
ERIC	7
IEEE Xplore	221
Science Direct	1,883
DOAJ	35

In order to ensure that we do not over or under-represent specific terminology, the output of these databases was compared. During this comparison, it was determined that Google Scholar represented the largest share of content, and that the other databases significantly duplicated content already present in Google Scholar. Due to this, we chose to isolate the rest of our methodology to only the content present on Google Scholar. While this does have the slight potential to bias the data, the sheer amount of content available via Google Scholar presents a significant corpus of work which should be representative of the current academic discourse. From here, relevant articles were downloaded for review via links as provided by Google Scholar. Where links were not available, original source materials were discovered via the use of Arxiv and other paper repositories, or were directly downloaded via the publishing bodies (e.g., university publishers, journal access, etc.). Using this downloaded data, we began to survey additional terminology co- associated with 'hallucination'. In order to validate the list of recorded co-associated terms, we utilized an 'AND' boolean search on Google Scholar. The string used was ' "Artificial Intelligence" AND "Hallucination" AND "term" '. This string allowed us to search across Google Scholar for incidences where the co-associated terms were used to describe a specific kind of hallucination, allowing us to replace the word

'term' with the observed terms in our downloaded documents. This resulted in a significant list of co-associated terms with high prevalence. These terms are shown below in Table II.

TABLE 2 Incidence of Co-Associated Terms ("Artificial Intelligence" AND "Hallucination" AND "term"

Error	12,000
Inference	11,800
Irrelevant	6,570
Inaccurate	5,540
Stochastic	4,780
Unrelated	3,350
Misinformation	3,310
Speculation	1,820
Epistemic	1,780
False Positive/False-Positive	1,440/1,440
Fabrication	1,380
False Negative/False-Negative	829/829
Misrepresentation	535
Confabulation	336
Overgeneralization/Over-Generalization	79/46
Undergeneralization/Under-Generalization	0/1

As an example for co-associated terms, Boussen et al. (2023). a paper submitted to the British Journal of Anesthesia, co-associated the term 'stochastic', meaning something which involves high random probability per Merriam-Webster (2024n), with 'hallucination', indicating a type of hallucination which arises from random generation rather than from inference or contextualization of the data set. In our data, this is listed as one of the results for the term 'stochastic'.

3 | RESULTS

The list of terms in Table II represent the most commonly co-associated terms within our data set as of December 2024. Based on these findings, we were able to group our findings into specific categories based upon similar factors or types of hallucination. To do this, we first defined the co-associated terms using the Merriam-Webster dictionary as follows:

- Error - "an act involving an unintentional deviation from truth or accuracy" - Merriam-Webster (2024c);
- Inference - "the act of passing from one proposition, statement, or judgment considered as true to another whose truth is believed to follow from that of the former"; "the act of passing from statistical sample data to generalizations (as of the value of population parameters) usually with calculated degrees of certainty" - Merriam-Webster (2024i);
- Irrelevant - "not relevant: INAPPLICABLE" - Merriam-Webster (2024j);
- Inaccurate - "not accurate: FAULTY" - Merriam-Webster (2024h);
- Stochastic - "involving a random variable"; "involving chance or probability" - Merriam-Webster (2024n);
- Unrelated - "not connected in any way: DISCRETE, SEPARATE" - Merriam-Webster (2024o);
- Misinformation - "incorrect or misleading information" - Merriam-Webster (2024k);
- Speculation - "an act or instance of speculating" - Merriam-Webster (2024m);
- Epistemic - "of or relating to knowledge or knowing" - Merriam-Webster (2024b);
- False Positive/False-Positive - "a person or test result that is incorrectly classified as positive [...] because of imperfect testing methods or procedures" - Merriam-Webster (2024f);
- Fabrication - "the act or process of fabricating" (Merriam-Webster, 2024), which is defined as "INVENT, CREATE" - Merriam-Webster (2024d); False Negative/False-Negative - "a person or test result that is incorrectly classified as negative [...] because of imperfect testing methods or procedures" - Merriam-Webster (2024e);
- Misrepresentation - "to give a false or misleading representation of usually with an intent to deceive or be unfair" - Merriam-Webster (2024l);
- Confabulation - "to fill in gaps in memory by fabrication" - Merriam-Webster (2024a); and
- Overgeneralization/Over-generalization and undergeneralization/Under-generalization - definition of generalization, as "a general statement, law, principle, or proposition" - Merriam-Webster (2024g);

These terms each fall into one of five categories.

The first of these categories is what we have termed Hallucination by Lacking Context. This category represents studies which co-associate hallucinatory behavior with generation which derives from lacking prompted context (e.g. not providing the specific form or function within the request itself) or from a lack of ability to parse the context within the data set (e.g. inability to identify specific elements of input material or training data as relevant to the prompt, failing to generate useful output or generating output which is accurate only insofar as the data set allows). Examples of this kind of hallucination include:

- The finding the GPT-vision, an image-driven LLM system, "is particularly sensitive to prompting, counterfactual text in images, and relative spatial relationship", as seen in Hwang et al. (2023);
- Comparative testing of advanced contextual prompting systems such as Retrieval Augmented Generation (RAG) which showed that "GPT-4 outputs with and without retrieval-augmented generation (RAG)" resulted in significant differences in accuracy, where "GPT-4 with RAG provided correct responses in 84% of cases" and "GPT-4 without RAG provided correct responses in only 57% of cases", per Ferber et al. (2024); and
- The "inherent constraints of these models in tackling mathematical and linguistic problem-solving tasks", as stated in Anderson et al. (2023), which makes contextual provision so important to generating valid output.

Compounding the complexity of this kind of hallucination is the fact that the lacking context may in some cases make the content generated seem more authentic than it actually is. In Kirstein et al. (2024), automated metrics for AI/LLM quality were found to insufficiently penalize specific kinds of errors, noting "count-based metrics predominantly show near-zero or negative correlations with errors, indicating they might not sufficiently penalize specific mistakes". Ultimately, this results in errors driven by lacking context which are duplicated and not sufficiently penalized, resulting in high speculation in output with little or no evidence backing it, as shown in McIntosh et al. (2023).

The next category of hallucination is Hallucination by Stochasticism. Stochasticism in the context of AI is the variable randomness inherent in any system Merriam-Webster (2024n), combined with both the highly complex nature of LLMs and their probabilistic sampling errors inherent in their data sets as seen in Watson and Cho (2024), resulting in random hallucination which "fills in the blanks" when no data is available.

In many cases, this type of error results from training the system upon recursive internal data sets or patterns - in other words, the LLM answers a question utilizing recursive randomness derived from past answers and errant data rather than objective truth or fact as seen in Li (2023). This random generation is actually a useful element in some cases, such as the use of LLMs to generate new artistic ideas or materials, as in Wang et al. (2024), but in the case of factual output, this randomness significantly reduces the quality and accuracy of generated output.

Thirdly, we identify Hallucination by Epistemology. This type of error is a computerized version of a philosophical quandry focused on the nature of knowledge. The study of knowledge and its nature is called epistemology - Greco (2017). Where this error arises in LLMs is in the inability for the system to differentiate between "knowing" something and "believing" in something.

While it is beyond the scope of this paper to decide whether or not an Artificial Intelligence has the capacity to "believe" in something, we can use the argument offered in Pollock (1984) that a reasonable belief can arise both from the "general reliability of a process" or the "single case reliability of an individual belief". In other words, an LLM may firmly 'believe' that what it is saying is true based upon the data it has available to justify it, even if that 'belief' is counter-factual.

Thus, it is possible for an LLM to 'believe' or 'know' something, and from this justified belief, to inaccurately respond to prompts. In fact, in many of the studies reviewed by this paper, a variety of phrases such as "ChatGPT Knows" (see Martínez et al. (2023), Apostolopoulos et al. (2023), Benzon (2023), Heck (2023), and Kleinpell (2023)) and "ChatGPT Believes" (see Wang et al. (2023), Yu et al. (2024), Sullivan-Paul (2023), **?**, and Wu et al. (2023)) have been used interchangeably. In a broader epistemological context, and with the context of knowing based on context and cultural influence as cited by Khine (2008), it must be stated that an error can arise from an LLM based upon justified beliefs which are generated based upon biases or other data set errors, resulting in hallucinations which are as valid as other human errors made due to epistemological pitfalls.

Fourthly, we identify Hallucination by Survivalism. While current standards suggest that AI and LLM systems are not alive in the traditional sense as argued by De Collibus (De Collibus), they nonetheless have been trained using methodologies arising from human evolution. For example, the use of reward-based token modeling as noted by MS et al. (2024) is heavily based on reward-based systems; these systems have been widely used in the treatment of adolescent ADHD and other neurodivergent conditions as shown in Korth (2006), showing a direct link between human nature and evolutionary incentive alongside the technology that we have developed.

This has resulted in LLMs being built upon reward systems which create a vested interest to output a response to a given prompt even when there is no data or context from which to generate a response. When faced with the reality of either giving a response which is invalid and getting a reward or not giving a response and thus not getting a reward, the LLM will prefer to generate the wrong answer as seen in Bai et al. (2024). This can then be compounded over time, resulting in the model reinforcing objectively incorrect context or evidence to maintain performance and output with a focus on acquiring rewards rather than a focus on creating accurate responses - in some cases, like those observed by Chan et al. (2024), this can have a feedback loop which amplifies false memories, statements, and even internal hallucinations.

Finally, we identify Hallucination by Intention. Large data sets are required to train LLMs, and as such, these data sets and the models which train upon them are particularly sensitive to data source poisoning. Work such as Fu et al. (2024)show that LLM data sets have several vectors that can be targeted, and work by Zhang et al. (2024) have shown with reasonable evidence that only .1% of the overall pre-training data set must be poisoned for three-quarters of attack vectors to persist from pre-training to post-training application. In other words, the intention of the creator of each data set may persist through to the eventual data generation, meaning that errors and hallucinations may arise from intentional data manipulation rather than an intrinsic error within the model itself. Worryingly, such data may be difficult to identify, as intentional data poisoning looks similar to lack of contextual data, as shown by Siciliano et al. (2024).

4 | DISCUSSION

We posit five particular types of hallucinations being present in the current academic discourse:

- Hallucination by Lacking Context - hallucinations arising from an incomplete data sources, leading to misrepresentations, factual contradictions with observable truth, and both false positives and false negatives;
- Hallucination by Stochasticism - hallucinations arising from the randomness inherent in the Large Language Model and its supportive infrastructure, especially when reinforced by recursive data training on outputs within the systems themselves;

- Hallucination by Epistemology - hallucinations that convert inference into factual statement through justified belief in counterfactual falsehoods or misinformation, lending authority to statements or outputs which are otherwise lacking evidence.
- Hallucination by Survivalism - hallucinations due to systems which prioritize an answer over the accuracy of that answer, especially when such a system is not self-correcting or introspective and instead incentivizes a reward loop.
- Hallucination by Intention - intentional poisoning of the LLM or its data set to generate false data.

This is an exclusive list of potential hallucination types - it is possible that further reviews will generate additional types of hallucination or that other types exist but are not significantly co-associated within the existing literature.

Some of these hallucination types can be mitigated through some specific solutions. The rise of Retrieval Augmented Generation, or RAG, has come to prominence in the API space in recent years, and solutions such as Self-Refinement through Feedback and Reasoning and Supervised Finetuning have made some headway towards improving overall accuracy and correctness as seen in Tonmoy et al. (2024). Both Hallucination by Lacking Context and Hallucination by Intention can be mitigated by particularly focusing on prompt engineering as demonstrated by White et al. (2023) and Xiao et al. (2024), though the impact of these solutions are somewhat limited by the quality of the data source and the implementation of that data by the LLM in question as shown in Amatriain (2024).

Other types of hallucination are more difficult to mitigate due to the nature of their cause. While Hallucination by Lacking Context can be mitigated by adding verifiable and high-quality context, Hallucination by Epistemology is harder to navigate as it is underpinned by a foundational and philosophical question that humanity has yet to fully grapple with - this can be seen in McCarthy (1981), where the fundamental nature of epistemology in relation to Artificial Intelligence is surveyed in depth. Specific prompt engineering to enforce internal validation and feedback loops can help to establish a cost for inaccuracy or a determinant need to be correct, but this can introduce additional issues with Hallucination by Survivalism which require more substantial structural design for reward systems to ensure accuracy and efficacy as seen in Wang (2024).

Finally, Hallucination by Intention is perhaps the most difficult of these hallucinatory types to deal with. While some solutions have been proposed, including a Human-in-the-loop model as proposed in Demartini et al. (2020) and an increased focus on media literacy and logical thinking as noted in Washington (2023), these approaches grapple with variability in determining veracity, logical connection, the nature of truth, and deeper issues concerning opinion and fact as noted in Onayinka et al. (2024).

Additional research is needed to identify methods and models to ensure veracity of training data as well as prevent iterative training upon data which may be poisoned, circularly ingested by the LLM itself, or otherwise impacted by its systemic use.

5 | CONCLUSION

While different terminology has been used by various studies, we find that hallucinations in Artificial Intelligence and Large Language Models fall into one of five categories:

- Hallucination by Lacking Context - hallucinations arising from an incomplete data sources, leading to misrepresentations, factual contradictions with observable truth, and both false positives and false negatives;
- Hallucination by Stochasticism - hallucinations arising from the randomness inherent in the Large Language Model and its supportive infrastructure, especially when reinforced by recursive data training on outputs within the systems themselves;
- Hallucination by Epistemology - hallucinations that convert inference into factual statement through justified belief in counterfactual falsehoods or misinformation, lending authority to statements or outputs which are otherwise lacking evidence.
- Hallucination by Survivalism - hallucinations due to systems which prioritize an answer over the accuracy of that answer, especially when such a system is not self-correcting or introspective and instead incentivizes a reward loop.
- Hallucination by Intention - intentional poisoning of the LLM or its data set to generate false data.

We recommend that further study and research be conducted to validate the nature of these hallucinations and to attempt to identify other types of hallucinations. Additionally, we suggest the use of this specific type terminology to specify what kind of hallucination is present when reported to reduce the complexity and opaque nature of this problem.

6 | DISCLOSURE

This work was produced without any financial support. The authors have no affiliations to disclose, and no relationships which might unduly influence the findings within this work.

7 | LIMITATIONS

As this study principally focuses on the co-association of English terms with hallucinatory behavior, it is limited by the language of choice. Co-associated terms may be different depending on the language. Our reliance on Google Scholar similarly has the potential to bias the data set. While we believe the impacts of both these limitations are negligible, they are nonetheless remarkable.

8 | ETHICS STATEMENT

This work abides by ethical guidelines for academic discourse and for research in the field of Artificial Intelligence. We have done our best to ensure that appropriate references and citations have been generated for all included content. We have not used any LLM or AI systems for the generation of this work. While we are unable to ensure that all sources cited in this work are entirely free from bias or ethical concerns, we have committed ourselves to due diligence on each cited work, and the authors believe that the specific work cited and their specific references in-text are free from undue bias which may influence the conclusions made herein.

9 | CITED WORKS

REFERENCES

Amatriain, X. (2024). Measuring and mitigating hallucinations in large language models: amultifaceted approach.

Anderson, N., A. McGowan, L. Galway, P. Hanna, M. Collins, and D. Cutting (2023). Implementing generative ai and large language models in education. In *2023 7th International Symposium on Innovative Approaches in Smart Technologies*, pp. 1–6. IEEE.

Apostolopoulos, I. D., M. Tzani, and S. I. Aznaouridis (2023). Chatgpt: ascertaining the self-evident. the use of ai in generating human knowledge. *arXiv preprint arXiv:2308.06373*.

Athaluri, S. A., S. V. Manthena, V. K. M. Kesapragada, V. Yarlagadda, T. Dave, and R. T. S. Duddumpudi (2023). Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through chatgpt references. *Cureus 15*(4).

Bai, Z., P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou (2024). Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Baker, S. and T. Kanade (2000). Hallucinating faces. In *Proceedings Fourth IEEE international conference on automatic face and gesture recognition Cat. No. PR00580*, pp. 83–88. IEEE.

Benzon, W. L. (2023). Chatgpt's ontological landscape. *Available at SSRN*.

Boussen, S., J.-B. Denis, P. Simeone, D. Lagier, N. Bruder, and L. Velly (2023). Chatgpt and the stochastic parrot: artificial intelligence in medical research. *British Journal of Anaesthesia 131*(4), e120–e121.

Chan, S., P. Pataranutaporn, A. Suri, W. Zulfikar, P. Maes, and E. F. Loftus (2024). Conversational ai powered by large language models amplifies false memories in witness interviews. *arXiv preprint arXiv:2408.04681*.

De Collibus, F. M. Are large language models" alive"?

Demartini, G., S. Mizzaro, and D. Spina (2020). Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities. *IEEE Data Eng. Bull. 43*(3), 65–74.

Ferber, D., I. C. Wiest, G. Wölflein, M. P. Ebert, G. Beutel, J.-N. Eckardt, D. Truhn, C. Springfeld, D. Jäger, and J. N. Kather (2024). Gpt-4 for information retrieval and comparison of medical oncology guidelines. *NEJM AI 1*(6), AIcs2300235.

Fu, T., M. Sharma, P. Torr, S. B. Cohen, D. Krueger, and F. Barez (2024). Poisonbench: Assessing large language model vulnerability to data poisoning. *arXiv preprint arXiv:2410.08811*.

Google (2018). *Hallucinations in Neural Machine Translation*. Google.

Greco, J. (2017). Introduction: What is epistemology?

Heck, T. G. (2023). What artificial intelligence knows about 70 kda heat shock proteins, and how we will face this chatgpt era. *Cell Stress and Chaperones 28*(3), 225–229.

Hwang, A., A. Head, and C. Callison-Burch (2023). Grounded intuition of gpt-vision's abilities with scientific images. *arXiv preprint arXiv:2311.02069*.

Khine, M. S. (2008). Knowing, knowledge and beliefs. *Epistemological Studies across diveres Cultures. Dordrecht*.

Kirstein, F., J. P. Wahle, T. Ruas, and B. Gipp (2024). What's under the hood: Investigating automatic metrics on meeting summarization. *arXiv preprint arXiv:2404.11124*.

Kleinpell, R. (2023). Artificial intelligence knows the value of nurse practitioners–why can't other humans?

Korth, T. M. (2006). *The impact of causal attributions on treatment choice acceptability for children with attention deficit hyperactivity disorder*. The University of Nebraska-Lincoln.

Li, Z. (2023). The dark side of chatgpt: Legal and ethical challenges from stochastic parrots and hallucination. *arXiv preprint arXiv:2304.14347*.

Martínez, G., J. Conde, P. Reviriego, E. Merino-Gómez, J. A. Hernández, and F. Lombardi (2023). How many words does chatgpt know? the answer is chatwords. *arXiv preprint arXiv:2309.16777*.

McCarthy, J. (1981). Epistemological problems of artificial intelligence. In *Readings in artificial intelligence*, pp. 459–465. Elsevier.

McIntosh, T. R., T. Liu, T. Susnjak, P. Watters, A. Ng, and M. N. Halgamuge (2023). A culturally sensitive test to evaluate nuanced gpt hallucination. *IEEE Transactions on Artificial Intelligence*.

Merriam-Webster (2024a). Confabulate. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024b). Epistemic. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024c). Error. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024d). Fabricate. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024e). False negative. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024f). False positive. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024g). Generalization. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024h). Inaccurate. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024i). Inference. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024j). Irrelevant. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024k). Misinformation. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024l). Misrepresent. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024m). Speculation. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024n). Stochastic. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Merriam-Webster (2024o). Unrelated. In *Merriam-Webster.com dictionary*. Merriam-Webster.

Mjolsness, E. D. (1986). *Neural networks, pattern recognition, and fingerprint hallucination*. Ph. D. thesis, California Institute of Technology.

Monteith, S., T. Glenn, J. R. Geddes, P. C. Whybrow, E. Achtyes, and M. Bauer (2024). Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry 224*(2), 33–35.

MS, A., J. VG, and D. PS (2024). Efficient hybrid inference for llms: Reward-based token modelling with selective cloud assistance. *arXiv preprint arXiv:2409.13757*.

Onayinka, T. S., J. K. Opele, L. B. Adewole, and C. I. Agbasimelo (2024). Ethical implications and policy frameworks for ai-driven solutions to combat misinformation in digital media. *UNIZIK*

Journal of Educational Research and Policy Studies 17(3).
Pollock, J. L. (1984). Reliability and justified belief. *Canadian Journal of Philosophy 14*(1), 103–114.
Rawte, V., A. Sheth, and A. Das (2023). A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.
Siciliano, F., L. Maiano, L. Papa, F. Baccini, I. Amerini, and F. Silvestri (2024). Adversarial data poisoning for fake news detection: How to make a model misclassify a target news without modifying it.
Sullivan-Paul, M. (2023). *How would ChatGPT vote in a federal election? A study exploring algorithmic political bias in artificial intelligence*. Ph. D. thesis, School of Public Policy, University of Tokyo.
Tait, J. I. (1982). Automatic summarising of English texts. Technical Report UCAM-CL-TR-47, University of Cambridge, Computer Laboratory.
Thaler, S. L. (1995). "virtual input" phenomena within the death of a simple pattern associator. *Neural Networks 8*(1), 55–65.
Tonmoy, S. M. T. I., S. M. M. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das (2024). A comprehensive survey of hallucination mitigation techniques in large language models.
Wang, B., X. Yue, and H. Sun (2023). Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. *arXiv preprint arXiv:2305.13160*.
Wang, F. (2024). Lighthouse: A survey of agi hallucination. *arXiv preprint arXiv:2401.06792*.
Wang, Z., A. Li, Z. Li, and X. Liu (2024). Genartist: Multimodal llm as an agent for unified image generation and editing. *arXiv preprint arXiv:2407.05600*.
Washington, J. (2023). Combating misinformation and fake news: The potential of ai and media literacy education. *Available at SSRN 4580385*.
Watson, W. and N. Cho (2024). Hallucibot: Is there no such thing as a bad question? *arXiv preprint arXiv:2404.12535*.
White, J., Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt.
Wu, Y., Z. Li, J. M. Zhang, M. Papadakis, M. Harman, and Y. Liu (2023). Large language models in fault localisation. *arXiv preprint arXiv:2308.15276*.
Xiao, W., Z. Huang, L. Gan, W. He, H. Li, Z. Yu, H. Jiang, F. Wu, and L. Zhu (2024). Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback.
Yu, X., L. Liu, X. Hu, J. W. Keung, J. Liu, and X. Xia (2024). Fight fire with fire: How much can we trust chatgpt on source code-related tasks? *arXiv preprint arXiv:2405.12641*.
Zhang, Y., J. Rando, I. Evtimov, J. Chi, E. M. Smith, N. Carlini, F. Tramèr, and D. Ippolito (2024). Persistent pre-training poisoning of llms. *arXiv preprint arXiv:2410.13722*.