Last Words

The Pitfalls of Defining Hallucination

Kees van Deemter
Utrecht University
Department of Information and
Computing Sciences
c.j.vandeemter@uu.nl

Despite impressive advances in Natural Language Generation (NLG) and Large Language Models (LLMs), researchers are still unclear about important aspects of NLG evaluation. To substantiate this claim, I examine current classifications of hallucination and omission in data-text NLG, and I propose a logic-based synthesis of these classfications. I conclude by highlighting some remaining limitations of all current thinking about hallucination and by discussing implications for LLMs.

1. Introduction: Evaluating the Veracity of a Text

When computers produce text, the quality of the texts is of paramount concern. For this reason, a substantial body of work across Natural Language Generation (NLG) focuses on evaluation of generated text. Evaluation can offer insight into various aspects of the quality of a generated text; indirectly, by looking at a range of generated texts, it can tell us how well a given NLG technique works.

A key family of quality criteria, for most text *genres* at least, centers around what might be called the veracity of the text. I will take the view that assessing the veracity of a text means assessing whether the text "speaks the truth, the whole truth, and nothing but the truth". Veracity is crucial: If a text is lacking in veracity, then any other virtues that it may possess—such as fluency and clarity, for example—can only amplify the risks that arise when the text is read (cf., Crothers, Japkowicz, and Viktor 2023), because a well-written falsehood is more likely to be taken seriously than a badly written one.

To obtain a clear perspective, I will focus on traditional NLG tasks first, whose aim is to convert a structured input into a text. I will argue that, even with such a relatively simple task, when we assess the veracity of a text, we do not quite know what we are doing. I will highlight some flaws in current analyses of hallucination, and I will offer a synthesis that does not suffer from these flaws. After that, I will argue that all current analyses still suffer from some important limitations, and I will discuss implications for more complicated tasks, as addressed by Large Language Models (LLMs).

Submission received: 6 October 2023; accepted for publication: 30 December 2023.

https://doi.org/10.1162/coli_a_00509

2. Existing Analyses of Veracity

When rule-based NLG systems are assessed, veracity is sometimes taken for granted; with the ascent of neural architectures, researchers in various areas of NLG have emphasized that this is often not justified (Vinyals and Le 2015; Koehn and Knowles 2017; Rohrbach et al. 2018; Maynez et al. 2020; Dušek and Kasner 2020; Raunak, Menezes, and Junczys-Dowmunt 2021). Let's look briefly at three attempts to analyze what the problem is and what forms it can take. Since these issues come into focus most sharply in relation to Data-text NLG—where the input to the generator is structured data, not text—we focus on that type of NLG; other types of NLG will be discussed in Section 5.

One clear-headed, but coarse-grained, attempt came from Dušek and Kasner (2020), who discussed data-text NLG systems whose input was a set of atomic statements and whose output was a verbalization of each of these statements.[1] For some specified individual x, the input could be $Type(x) = Restaurant \wedge Food(x) = Italian \wedge Price(x) = Low$, and the output "$x$ is an affordable Italian restaurant". The authors highlighted two possible problems, namely, hallucination and omission. For them, **hallucination** occurs when the output does not follow from the input (so it's possible that the input is true but the output is false); **omission** happens when the input does not follow from the output (so it's possible that the output is true but the input is false). Thus, if the generator produces "x is an affordable *veggie* Italian restaurant", then the information that $Style(x) = vegetarian$ is hallucinated; if it produces "x is an Italian restaurant", then that involves an *omission*, because the information that $Price(x) = Low$ is omitted. Note that, by relying on the "follows from" relation, Dusek and Kasner's is, at heart, a *logical* analysis, where the input is a formal meaning representation formula and the output is natural language text. The authors phrase their idea in terms of Natural Language Inference (NLI), assuming a version of NLI that "crosses formats", relating formulas and text (see Section 4).

A more finegrained analysis of hallucination was offered by Ji et al. (2023), who distinguished between **intrinsic hallucination** and **extrinsic hallucination**: The former is "output that contradicts the source"; the latter is "output that can neither be supported nor contradicted by the source". This analysis, whose core concepts (contradict, supported) appear likewise, though more implicitly than in the previous case, to be rooted in logic, was applied to a number of NLG areas, including data-text NLG; notably, omission of information was not discussed.

A third strand of research tries to classify all kinds of errors that are found in computer-generated texts (Thomson and Reiter 2020; Moramarco et al. 2022; Van Miltenburg et al. 2023). Because these errors include hallucination, it will be instructive to examine one such account, namely, Thomson and Reiter (2020), who distinguished between outputs that contain an incorrect number, an incorrect named entity, an incorrect word (that is neither an incorrect number nor an incorrectly named entity), "non-checkable" information, context errors, and other kinds of errors.

To see how these three analyses compare, let's look at some outputs that may be generated on the basis of a given input. Suppose the input is, once again, $Type(x) = Restaurant \wedge Food(x) = Italian \wedge Price(x) = Low$.

1 In other words, the decision of "What to say" (as opposed to "How to say it"), also known as Content Determination, has already taken place (see, e.g., Gatt and Krahmer 2018, Section 3).

Ex1. "*x* is an Italian restaurant."
D&K: Omission. **Ji** : n.a. **T & R**: n.a.

Ex2. "*x* is an affordable veggie Italian restaurant".
D& K: Hallucination. **Ji** : Extrinsic hall. **T&R**: word error.

Ex3. "*x* is a veggie restaurant".[2]
D&K: Hallucination and Omission. **Ji** : Extrinsic hall. **T&R**: word error

Ex4. "*x* is an affordable Norwegian restaurant".
D&K: Hallucination and Omission. **Ji** : Intrinsic hallucination. **T&R**: n.a.

Similar patterns emerge from different kinds of input. Suppose, for example, a numerical input for weather reporting specifies that the temperature is above 22° Celsius. Now consider the output "The temperature is above 21 degrees". With this type of input, the definitions are not entirely clear cut (e.g., because it can be difficult to isolate the precise part of the output that is at fault), but I take it that both **D&K** and **Ji** would analyze this output exactly like (Ex1) above; "The temperature is above 23" would be analyzed like (Ex2); "The temperature is below 25" would be analyzed as Ex3, and "The temperature is below 22" as Ex4. I take it that **T&R** would analyze each of these outputs as an Incorrect Number error.

Clearly, these analyses differ substantially from each other, echoing the complaint in Huidrom and Belz (2022) that classifications of NLG errors tend to disagree with each other. Moreover, each analysis conflates two or more types of misinformation, each of which would pose different kinds of risks if they occurred in real life. Dusek and Krasner's analysis, for example, does not distinguish between (Ex3) and (Ex4), because all it can say about both situations is that omission and hallucination occur (because the input does not imply the output and the output does not imply the input). Ji et al. conflate (Ex2) and (Ex3), because both "can neither be supported nor contradicted by the source".

If we wish to use concepts such as hallucination as a tool for evaluating the veracity of NLG models (using either human annotations or computational metrics or both), and as a starting point for mitigating different types of generation errors, then we need to go back to the drawing board.

3. A Synthesis of Existing Analyses

Luckily, a synthesis of the analyses by Dusek and Kasner and Ji et al. is possible. To obtain a systematic perspective that is applicable to domains of all kinds, let us step back and ask what Logical Consequence (i.e., "follows from", "$\models$") relations can exist between input and output, assuming a classical logic. Assume, for now, that neither the input nor the output to the generator is internally inconsistent. It can be true or false that *input* $\models$ *output*; likewise, it can be true or false that *output* $\models$ *input*; furthermore, if

2 Even with this simple input, it is not always obvious how the proposed definitions should be applied. For example, Thomson and Reiter's scheme might alternatively be used to categorize Ex2 as non-checkable, and Ex4 as involving an incorrectly named entity (namely, Norway).

input $\not\models$ *output* and *output* $\not\models$ *input*, it matters whether *input* $\models \neg$*output* (as in Ji et al.'s intrinsic hallucination, where output contradicts input), splitting error type 3 in two:[3]

0. *input* $\models$ *output* and *output* $\models$ *input*. (Input and output are well matched.)
1. *input* $\models$ *output* but *output* $\not\models$ *input*. (Output is too weak.)
2. *input* $\not\models$ *output* but *output* $\models$ *input*. (Output is too strong.)
3. *input* $\not\models$ *output* and *output* $\not\models$ *input*. (Neither follows from the other.)

    3a. (3) and *input* $\not\models \neg$*output*. (Input and output are logically independent of each other). E.g., "*x* is a veggie restaurant."

    3b. (3) and *input* $\models \neg$*output*. (Input and output contradict each other.)
    E.g., "*x* is an affordable Norwegian restaurant."

If one also wishes to take into account that outputs can be self-contradictory (as documented by Moramarco et al. 2022, for example) or tautologous, the logical analysis can be pushed further by splitting category 1 and 2 as well:[4]

1. *input* $\models$ *output* but *output* $\not\models$ *input*. (Output is too weak.)

    1a. (1) and $\not\models$ *output*. (Output is not tautologous; this is the normal case). E.g., "*x* is an Italian restaurant."

    1b. (1) and $\models$ *output*. (Output is tautologous). E.g., "*x* is Italian or not."

2. *input* $\not\models$ *output* but *output* $\models$ *input*. (Output is too strong.)

    2a. (2) and $\not\models \neg$*output*. (Output is not contradictory; this is the normal case). E.g., "*x* is an affordable veggie Italian restaurant."

    2b. (2) and $\models \neg$*output*. (Output is contradictory).
    E.g., "*x* is a veggie steak restaurant."

It is easy to see how this new analysis applies to other types of information, for example, when the input is numerical. For instance, if the input says the temperature is between 20° and 30° Celsius, then an output that says "The temperature is between 25° and 35°" would be logically independent of the input and hence inhabit category 3a.

Note that, for better or worse, the *truth* of the output, in the real world, has not been a consideration: our core question has been whether the output of the generator matches the input.

3 In (1), *input* $\models \neg$*output* would imply that the *input* is inconsistent (because also *input* $\models$ *output*). In (2), *output* $\models \neg$*input* would imply that the *output* is inconsistent (because also *output* $\models$ *input*).

4 In classical logic: a tautology follows from everything (1b); everything follows from a contradiction (2b).

4. Limitations of These Analyses

This synthesis can underpin a computational metric that records how often each hallucination type occurs in a given (generated) text, provided the "follows from" relation can be computationally modeled using Natural Language Inference (NLI) (as proposed by Dušek and Kasner 2020). This presupposes that typos and incorrect names should somehow be finessed (e.g., Faille, Gatt, and Gardent 2021), which may or may not be seen as affecting the veracity of a text. More problematically, it assumes that NLI is able to deal with ambiguous and vague text. For instance, if the input specifies that the temperature is above 22° Celsius, does it follow that "It's a warm day"? Such issues may well be beyond the present state of the art of NLI (Liu et al. 2023).

Our analysis has some intrinsic limitations. It's best seen as the topic-independent part of a full analysis, which disregards everything that's specific to a specific domain or application. When applied to an NLG system whose texts offer treatment advice to doctors, for example, then further distinctions will need to be made to determine whether an error is medically significant or not (cf., Moramarco et al. 2022).

Pragmatic Reasoning Should Always Be Applied. When a multi-sentence text is generated, each sentence should be given an interpretation appropriate for its context,[5] for instance with anaphoric pronouns resolved. Likewise, any pragmatic implicatures (in the style of Grice 1975) of the sentence should be taken into account. Similar remarks apply to presupposition, irony, and metaphor, which go beyond what is stated literally in a text, yet all of which can cause hallucination and omission. For example, when the weather report for a sunny day speaks metaphorically of "wall-to-wall sunshine", then a narrowly semantic analysis might misclassify this as a (3a-type) hallucination (after all, the input does not mention any walls), but an analysis that understands metaphor should consider it to be truthful (i.e., well matched). In other words, measures should be taken to ensure that the "follows from" relation is pragmatically, as well as semantically, aware.

Even a Domain-independent Analysis Could Be Driven Further. The classical "follows from" relation cannot tell us everything one might want to know about the relation between input and output. For a start, it does not tell us anything about the *amount* of information that is added to or omitted from the input, only whether there exists such information. More subtly, suppose our input is, once again, $\mathit{Type}(x) = \mathit{Restaurant} \wedge \mathit{Food}(x) = \mathit{Italian} \wedge \mathit{Price}(x) = \mathit{Low}$. Then the output "$x$ is a Norwegian restaurant" falls into the same category as "x is an affordable Norwegian restaurant" even though, unlike the latter, this output *omits* some information ("affordable") from the output. And since our analysis is based on classical logic, it is unable to tell us what a text is *about*. Assuming the same input as before, an output that says "The cat is on the mat" would fall into the same category as "x is a veggie restaurant", because both outputs are logically independent of the input; the fact that the sentence about the cat is also topically unrelated to the input is something that our analysis is unable to pick up. I take it that these limitations are acceptable once we are aware of their existence.

To assess whether an NLI-based implementation of our analysis matches the opinions of human judges, we are currently conducting a pilot study in which domain experts are asked to apply our analysis to commercial advertisements. A challenge facing us, in this domain, is a situation in which the NLG output contains information that is not present in the input, but where the added information can nonetheless

5 This echoes the idea of Context Errors, a separate category in Thomson and Reiter (2020) .

be inferred with high probability. The question is, are these hallucinations or not? In other cases, the texts venture gratuitous information like "This place has an amazing atmosphere". Although such information cannot be inferred from the input, we are encouraging annotators to turn a blind eye because the falsity of the added information would be difficult to prove in a court of law (e.g., if a customer decided to complain about the atmosphere at this place). Arguably, this approach goes beyond the proposal in Section 3, because it asks not whether the output follows from the input, but whether the output is (or is likely to be) true in the real world. It is time that we examine this distinction more closely.

5. Veracity in Large Language Models and in Other Tasks than Data-text NLG

With the current popularity of LLM-driven Generative AI, the question is being raised, not only by researchers but by society at large, what is the veracity of the texts that are generated by LLMs. These issues matter greatly because LLMs are starting to be employed in real-world contexts, by people who are no experts in NLP, and who may not always be aware how veracity may be compromised. Identifying the ways in which LLMs can go wrong is a necessary first step towards identifying real-world risks and technical mitigation strategies.

Accordingly, a number of research groups have started to investigate how veracity problems in LLMs can be classified and mitigated (Zhang et al. 2023; Huang et al. 2023), discussing a wide range of interesting problems and potential solutions. The question comes up whether, on top of these important endeavors, it might be possible to arrive at a more systematic analysis along the lines of Section 3, which might reduce the risk that error categories might be overlooked or unclearly defined.

So, what happens when one tries to define the different ways in which LLMs can be lacking in veracity? LLMs have been used to generate texts for a large variety of purposes. These differences are so substantial that veracity cannot mean the same thing in all these cases. In fact, for purposes of error classification, it does not matter what architecture (e.g., using LLMs or some other technique) is utilized to perform a certain NLP task. But, contrary to what some writing in this area might lead one to expect, it matters hugely *what that task is*; let's see why.

The analyses of veracity discussed in Section 3 centered around the relationship between the input of the NLG system and its textual output, asking whether the output is too weak, or too strong, for its input, for instance. This perspective carries over almost verbatim when the purpose of an LLM is to express all the information in the input truthfully, as in Lorandi and Belz (2023) or Yuan and Färber (2023), for example. (Since LLMs operate on text, their generation process starts by converting structured input into a linearized format, using mark-up to convey logical structure [cf., Harkous, Groves, and Saffari 2020], but I take this conversion step to be almost trivial.) Consequently, all the distinctions that were made in Section 3 apply to these LLM uses as well, including distinctions that are seldom made in the literature on LLMs, such as the distinction between 1a and 1b, and the one between 3a and 3b. The perspective of Section 3 may also carry over to some types of Table-to-Text generation and Question Answering; it might also apply to Machine Translation (e.g., Dale et al. 2022), which likewise hinges on an equivalence between input and output; the main difference with the type of classical NLG discussed in earlier sections is that where the latter generate text from structured input, the former generate text from text.

Some other uses of LLMs take up a middle ground, where the perspective of Section 3 is partially, but not wholly, applicable. When LLMs (or other NLP architectures,

for that matter) are used for text summarization (e.g., Pu, Gao, and Wan 2023) or caption generation (e.g., Rotstein et al. 2024), the requirement that *input* $\models$ *output* still makes sense (because a summary is meant to be faithful to the thing that is summarized), but the requirement that *output* $\models$ *input* requires nontrivial modification (because only the most important aspects on the input need to end up in the output summary). Our analysis is applicable in full if and only if the task of the NLG system is to express all information in its input.

Defining hallucination is even more problematic in open-ended LLM applications such as unrestricted Question Answering (see Bubeck et al. 2023; Zhang et al. 2023; Huang et al. 2023 for plenty of examples) or essay writing (e.g., Fitria 2023) because, in such applications, it would be difficult to say *what input* (in our sense of that term) such an LLM is expressing.[6] As was pointed out by Zhang et al. (2023), such systems suffer not only from what these authors call "input-conflicting" information (where the LLM's output deviates from the prompt provided by users), and from "context-conflicting" information (where the output conflicts with information previously generated by the system), but also, most important of all, from "fact-conflicting" information (where the output conflicts with world knowledge). Note that the output of such LLMs can still be assessed in terms of whether it is tautologous (our type 1b) and whether it is contradictory (type 2b); testing for other kinds of fact-conflicting information, however, would require assessing whether the output is objectively true (replacing our earlier question of whether *input* $\models$ *output*), and possibly also whether the output is informative enough for the application at hand (replacing the question of whether *output* $\models$ *input*); in practice, of course, such assessments are sometimes difficult to make with certainty. In a political essay, for example, who is to say what is objectively true, or what it means for an essay to be informative enough? The idea that one can always answer such questions objectively is optimistic to say the least.

But although this is treacherous territory, formal logic might once again help us on our way, via a logic of beliefs, desires, and intentions (BDI) (Bratman 1987). (See also van Ditmarsch, Hendriks, and Verbrugge [2020] for a broader inter-disciplinary perspective.) BDI logic adds to our analytical arsenal because it allows us to reason about the inferences that a hearer will draw from an utterance, and to formally elucidate notions such as lying and misleading (Sakama, Caminada, and Herzig 2010). Suppose, for example, today's weather forecast does not mention a hurricane, then listeners may infer that no hurricane will take place; if a hurricane hits them nonetheless, they were misinformed by the forecast.[7] Omissions of this kind are, in the terminology of Sakama and colleagues, "withholding information"; they are important because misinformation often hinges on strategic omissions. Another type of misleading highlighted by Sakama and colleagues is "half truth": An output is a half truth if a false proposition r is not communicated directly, yet a truthful output is generated of which the hearer believes that r follows from it.[8] Sakama et al.'s example is of a speaker bragging that he holds a permanent position at a company, without saying that the company is almost bankrupt.

6 If one wishes to view all the data that determine the LLM's response as input to the generator, this implies that such an input tends to be internally inconsistent, necessitating the use of a para-consistent, non-classical logic (see, e.g., Priest, Tanaka, and Weber 2022). Of course the use of a paraconsistent logic would complicate the idea of seeing the output of a generator as "following from" its input considerably.

7 Using q to abbreviate "There will be a hurricane", and $C_{SH}p$ to say that the speaker (S) communicates to the hearer (H) that p, and using $B_{H}p$ to say that the Hearer (H) believes that p, this can be expressed (slightly modifying the formalism of Sakama, Caminada, and Herzig [2010]) as the conjunction $\neg C_{SH}q$ *and* $B_H(q \rightarrow C_{SH}q)$ *and* $\models q$.

8 In modified BDI notation, with p as the output, $C_{SH}p$ *and* $\neg C_{SH}r$ *and* $B_H(p \rightarrow r)$ *and* $\models p$ and $\not\models r$.

A BDI analysis does not by itself tell us whether a given LLM-generated text withholds information, or tells a half truth; no existing NLI system would be able to tell us whether this is the case. On the other hand, BDI logic may give us a starting point for understanding these matters. It would be interesting, for example, to try out whether annotators might be able to apply a version of Sakama, Caminada, and Herzig's (2010) categories to essay texts; in this case, substantial levels of disagreement between annotators are to be expected.

6. Conclusion

Natural Language Processing has made great strides, but our lack of understanding of some of the most important issues in the evaluation of generated text should be cause for humility, and even worry. To start addressing this problem, the NLP community should be prepared to do two unfashionable things: First, it should liaise with logicians; after all, the latter have always focused on concepts like truth and evidence. The synthesis offered in Section 3, and the more speculative ideas in Section 5, indicate the benefits that can be gained in this way. Secondly, work on NLG evaluation should pay more attention to the most difficult aspects of communication, including phenomena such as ambiguity and vagueness, which undermine our current understanding of logical inference (see e.g., Van Deemter 2010), and including the crucial distinction between what a sentence asserts and what an utterance of the sentence communicates *via* such mechanisms as implicature, presupposition, irony, and metaphor (see, e.g., Levinson 1983).

Now that all of us are becoming consumers of computer-generated text, the NLP community disregards such matters at everyone's peril.

Acknowledgments

I am grateful for comments from reviewers; from people in Utrecht's NLP group; and from Lasha Abzianidze, Guanyi Chen, Hans van Ditmarsch, and Ehud Reiter.

References

Bratman, Michael. 1987. *Intention, Plans, and Practical Reason*. University of Chicago Press.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Crothers, Evan, Nathalie Japkowicz, and Herna L. Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002. https://doi.org/10.1109/ACCESS.2023.3294090

Dale, David, Elena Voita, Loïc Barrault, and Marta R. Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. *arXiv preprint arXiv:2212.08597*. https://doi.org/10.18653/v1/2023.acl-long.3

Dušek, Ondřej and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation (INLG)*, pages 131–137. https://doi.org/10.18653/v1/2020.inlg-1.19

Faille, Juliette, Albert Gatt, and Claire Gardent. 2021. Entity-based semantic adequacy for data-to-text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1530–1540. https://doi.org/10.18653/v1/2021.findings-emnlp.132

Fitria, Tira Nur. 2023. Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. In *ELT Forum: Journal of English Language Teaching*, 12:44–58. https://doi.org/10.15294/elt.v12i1.64069

Gatt, Albert and Emiel Krahmer. 2018. Survey of the state of the art in natural

language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170. https://doi.org/10.1613/jair.5477

Grice, Herbert P. 1975. Logic and conversation. In *Speech Acts*. Brill, pages 41–58. https://doi.org/10.1163/9789004368811_003

Harkous, Hamza, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! End-to-end neural data-to-text generation with semantic fidelity. *arXiv preprint arXiv:2004.06577*. https://doi.org/10.18653/v1/2020.coling-main.218

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Huidrom, Rudali and Anja Belz. 2022. A survey of recent error annotation schemes for automatically generated text. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 383–398. https://doi.org/10.18653/v1/2022.gem-1.33

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38. https://doi.org/10.1145/3571730

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*. https://doi.org/10.18653/v1/W17-3204

Levinson, Stephen C. 1983. *Pragmatics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511813313

Liu, Alisa, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. *arXiv preprint arXiv:2304.14399*. https://doi.org/10.18653/v1/2023.emnlp-main.51

Lorandi, Michela and Anya Belz. 2023. Data-to-text generation for severely under-resourced languages with GPT-3.5: A bit of help needed from Google Translate. *arXiv preprint arXiv:2308.09957*.

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*. https://doi.org/10.18653/v1/2020.acl-main.173

Moramarco, Francesco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. *arXiv preprint arXiv:2204.00447*. https://doi.org/10.18653/v1/2022.acl-long.394

Priest, Graham, Koji Tanaka, and Zach Weber. 2022. Paraconsistent logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2022 edition. Metaphysics Research Lab, Stanford University.

Pu, Xiao, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Raunak, Vikas, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. *arXiv preprint arXiv:2104.06683*. https://doi.org/10.18653/v1/2021.naacl-main.92

Rohrbach, Anna, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*. https://doi.org/10.18653/v1/D18-1437

Rotstein, Noam, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. 2024. FuseCap: Leveraging large language models for enriched fused image captions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5689–5700.

Sakama, Chiaki, Martin Caminada, and Andreas Herzig. 2010. A logical account of lying. In *Logics in Artificial Intelligence: 12th European Conference, JELIA 2010, Helsinki, Finland, September 13–15, 2010. Proceedings 12*, pages 286–299. https://doi.org/10.1007/978-3-642-15675-5_25

Thomson, Craig and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168. https://doi.org/10.18653/v1/2020.inlg-1.22

Van Deemter, Kees. 2010. *Not Exactly: In Praise of Vagueness*. Oxford University Press.

van Ditmarsch, Hans, Petra Hendriks, and Rineke Verbrugge. 2020. Editors'

review and introduction: Lying in logic, language, and cognition. *Topics in Cognitive Science*, 12(2):466–484. https://doi.org/10.1111/tops.12492, PubMed: 32118362

Van Miltenburg, Emiel, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Stephanie Schoch, Craig Thomson, and Luou Wen. 2023. Barriers and enabling factors for error analysis in NLG research. *Northern European Journal of Language Technology*, 11. https://doi.org/10.3384/nejlt.2000-1533.2023.4529

Vinyals, Oriol and Quoc Le. 2015. A neural conversational model. *CML Deep Learning Workshop*.

Yuan, Shuzhou and Michael Färber. 2023. Evaluating generative models for graph-to-text generation. *arXiv preprint arXiv:2307.14712*. https://doi.org/10.26615/978-954-452-092-2_133

Zhang, Yue, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.