Self consistency and beyond - in the search of decoding method for mitigating hallucinations

Noa Mark
Tel-Aviv University
noamark@mail.tau.ac.il

Yarden Frenkel
Tel-Aviv University
yf2@mail.tau.ac.il

Abstract

State of the art models in abstractive summary frequently contain hallucinations (> 30%). Defining and evaluating hallucinations is an open challenge. There is no state-of-the-art metric for evaluation, and no clear protocol or method for mitigation. Recent work has been done in improving LLM generation through self-consistency. We wish to evaluate this method as a tool for mitigating hallucination, as well as suggesting our own decoding-based strategy to reduce factual inconsistency. Our model leverages LLM paraphrasing functionality to create multiple versions of the input document, thereby creating diversity in the output signal. This new way to create a self-ensemble model is aligned with the idea that summaries should be robust to paraphrasing. We found that both self-consistency and paraphrasing significantly improve the percentage of factual errors and extrinsic hallucination in the summaries. Additionally, there is no clear advantage to paraphrasing.

1 Introduction

Different approaches exist to mitigate hallucinations (Ji et al., 2023). We can categorize them into data-related methods and modeling and inference methods. Data related methods include building better and faithful datasets, cleaning data automatically and information augmentation. Modeling methods include careful architecture design, adjust training method and multi-task learning. In this work we are focusing on inference mitigation methods. Which allows us to leverage state-of-the-art models with small changes in the generation pipeline.

Self-consistency (Wang et al., 2023) work is a great example for inference method/decoding strategy. The key concept is generating multiple outputs using different seeds and selecting the most-consistent answer. They show this model outperforms other ensemble-based and sampling methods. The idea underlying self-consistency is that complex reasoning tasks often offer multiple viable paths for reaching the correct answer through reasoning. The truth is consistent, so as long as our errors are not too-correlated, and we can sample diverse answers, we can hope that the majority answer will be closer to the truth. Avoiding heavy tailed distribution of errors. Due the original paper was limited to close-domain tasks, it was recently expanded to open-generation problems (Jain et al., 2023).

This line of work was not yet explored in this context of hallucination. As we know from previous evaluation work (Pagnoni et al., 2021, Kryściński et al., 2019, Ji et al., 2023), improving the model ROUGE score or other performance metrics, does not indicate reducing the hallucination or the factual errors in the model. Sometimes, it's even the opposite, as the expressiveness power of the model grows and so does the chance for hallucination. Nevertheless, following the original paper's intuition that the truth is consistent, we hoped that this method will improve factual-consistency of the model and reduce some types of hallucinations.

In this work [1] we will examine whether self-consistency is a good method for improving factual-consistency, and suggest and evaluate another inference method to mitigate hallucination. We chose to focus on hallucination in the task of abstractive-summaries. As it's currently extensively researched in the context of hallucination (Ji et al., 2023) both in defining the meaning and types of hallucinations, ways to evaluate them and methods to reduce the likeliness they appear. Specifically, in an abstractive summary, we have an input document and a short abstractive summary. Therefore, there could be both factual errors, as well as information that might be true but does not appear in the reference text. In other words, we look for both intrinsic

[1] Our code: https://github.com/yardenfren1996/SelfConsistency

and extrinsic hallucinations (as defined in Ji et al. 2023), and aim for factual consistent summaries.

2 Proposed method

2.1 Definition

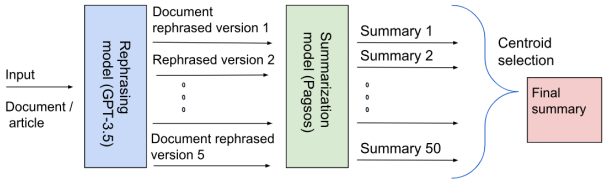

Figure 1: Given an input, make the model rephrase it, answer to each of the different phrases. The "Truth" should be consistent, and appear frequently. So we could choose to aggregate the final answer based on the different phrases results. This is a post-processing method for dealing with hallucinations.

The self-consistency method achieves diversity in its output through the utilization of various seeds. In this study, we aim to assess consistency using a different approach. Instead of employing different seeds, we employ a paraphrasing model to rephrase the inputs. This should enforce the output to be robust to different paraphrasing, while focusing on the meaning. Similar to the training-augmentation technique built in Kryściński et al., 2019. It is also aligned with our common sense, as summarization is often abstracting a document in different words.

2.2 Challenges

Creating paraphrase versions of the input document is a challenging task. It requires handling of long-sequences. We tried to use LLama-2 (Touvron et al., 2023) for this task, but we could not run it on the university resources, even when using 8-bit quantization (Jacob et al., 2017). Later, we tried to use other open-source LLM based for this task, such as Vladimir Vorobev (2023) - they raise the challenge of overcoming their limited input length. We tried to rephrase single sentences using them, thinking that we could create a version of the document combining these paraphrased sentences. Overcoming the incoherent that might accrue when combining single sentences into one document by trusting the power of the crowd using our centroid selection mechanism. But, this model showed poor performance in our manual evaluation even on single sentences: omitting important details, interrupting sentences, etc. Finally we resort to commercial products. We consider using the AI21 model specialized in paraphrasing. But this model is also limited to only a few sentences - up to 500 characters - while a document can have thousands of characters. So, this solution still leaves us with the problem of combining the sentences. Finally, we manually compared the chat version of both j2 (light, med) by AI21 and GPT by OpenAI (OpenAI, 2023), used with zero-shot for paraphrasing. The light version of J2 results in severe hallucination in paraphrasing. While GPT-3.5 created diverse and coherent paraphrasing of the whole document. Therefore, we decided to use GPT by OpenAI to create paraphrased inputs. Because of its cost, both financially and timely (there is limitation on request frequency in the API) we used small random-sample 1% of the test sets, aka 117 samples, for this evaluation.

3 Experiments

3.1 Defining the centroid

Both paraphrasing the input-document and self-consistency method worked under the assumption that we can define a reasonable centroid. Wang et al., 2023 defines the centroid by selecting the most common answer as it only applied on close-domain questions. Jain et al., 2023 defined the centroid using simple consistency-scores based on n-gram or weighted n-gram.

$$GSC_{sim}(i) := \frac{1}{M-1} \sum_{j,j \neq i} sim(i,j)$$

$$sim_{ngram}(i,j) := \frac{1}{|V|} v_i \cdot v_j$$

Jain et al., 2023 mentioned limited work using OpenAI embedding. They chose not to work with embedded based scoring, because it's heavy, and for their main task, coding, it did not outperform. Although OpenAI embedding based scoring did seem to outperform for summaries (see Appendix F). Another downgrade, not mentioned in their work, is that OpenAI is not open-source and weirdly used as BERT. In our work, we further investigate the improvement achieved by using open-source embedding models for the similarity functions.

Using embedding models such as BERT (Devlin et al., 2019) or GTE (Li et al., 2023) Consistency scores are defined the same as GSC above, where the similarity function is the sum over cosine similarity between the embedded-tokens of the two possible summaries.

	Embedded-based score			ROUGE-2 score		
	BERT centroid	GTE centroid	N-gram centroid	BERT centroid	GTE centroid	N-gram centroid
Base model	77	71	21	16	16	16
Self-consistency	81	93	34	22.45	22.5	22.5
Paraphrasing	80.3	92.8	32	20.5	20.8	20.7

Table 1: Paraphrasing vs. self-consistency

Comparing the performance of BERT embedding based score to N-Gram based score in an abstractive summary task, we can see that the embedding model leads to improvement in selection of the centroid. As BERT is widely used both for modeling and evaluating summaries performance - the following experiments are done using BERT scores.

3.2 Paraphrasing vs. self-consistency

We compared paraphrasing and self-consistency mitigation methods using Pegasus (Zhang et al., 2019) SOTA for summarization task on the XSUM dataset (Narayan et al., 2018). For evaluation we looked at the Rouge-2 metric, as well as, for both BERT (Devlin et al., 2019) and GTE (Li et al., 2023) we compared the centroid selected by their embedding model to the ground truth using the similarity metric defined by this model (See Table 1) . We get that both self-consistency and paraphrasing methods outperform the original model across all metrics. But self-consistency slightly outperforms paraphrasing. It is important to note, due to limited scope and resources, we only looked at specific hyper-parameters settings, and we did not examine the effect of different numbers of paraphrasing/seeds and beam-search on paraphrasing performance. We limited ourselves to a setting where the number of paraphrased versions equal the number of seeds. This result was achieved by 300 samples (2.5%) randomly selected from the test set.

3.3 Mitigating hallucinations

As we know from previous work (Pagnoni et al., 2021, Kryściński et al., 2019, Ji et al., 2023), the typical metrics used to evaluate summarization algorithms do not consider whether the summaries are factually accurate in relation to the source documents. Following Pagnoni et al., 2021 survey we know that currently the state-of-the-art evaluation tools for hallucination are not aligned with human

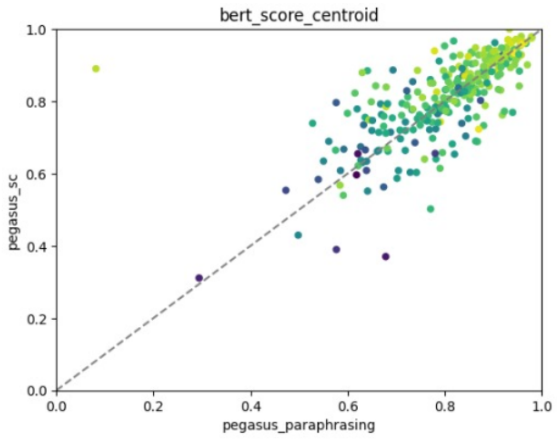

Figure 2: Compare self-consistency vs. paraphrasing performance according to BERT-score. Each dot is a sample, overall 300 (2.5%) randomly selected from the test set. The samples are colored according to the performance of the original model (without any intervention). We can see the performance is quite correlated, both between the intervention method, and between the intervention method to the original model. Aka those mitigation methods don't seem to be complementary to one another or help to mitigate different kinds of problems. While self-consistency slightly outperforms paraphrasing according to the conventional scores (Both Embedding based and N-gram based).

evaluation. Therefore, to answer our question, are self consistency and paraphrasing good methods for mitigation hallucination we manually tagged 1% of test samples. We compared each of the suggested summarization, of the original method, the self-consistency and the paraphrasing-mitigation, against the source document and tagged the type-of-error that exists following Pagnoni et al., 2021 topology of factual-errors (See Appendix). As we know that the quality of the dataset may account for hallucinations as well, we also added a flag of ambiguity of the target summary with respect to the input article (See Table 2 for results).

As reflected in the result in Table 2, Self-consistency and paraphrasing both help to reduce hallucinations. Self-consistency helps to reduce

	Total	Original model	Self-consistency	Paraphrasing
# Numbers	117	63	50	49
% Percentage	100%	53%	42 %	41%

Table 2: Hallucination count out of the manually tagged 1% samples of test sets, given different mitigation method.

factual-errors in 25% (16 samples). This result is statistically significant, and the null hypothesis that the % of the errors of the original and self-consistency models, was drawn from the same distribution is rejected with p-value « 0.0033 (see figure 3). Currently, no significant difference has been found between the methods.

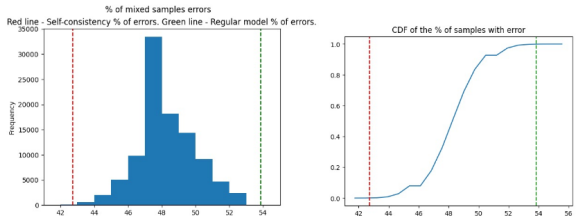

Figure 3: We used a permutation test to know whether the result that the % errors using self-consistency is lower than the % of errors in the original model is statistically significant. The null hypothesis is that no model is better than the other. Then, we will expect the % of errors in samples drawn from both models, aka sometimes using the regular model and sometimes the self consistency model, will have the same distribution as each of them. We call this 'mixed sample errors' and plot its pdf and cdf above. Note, we are looking at all the documents, and switching the model being used for inference. So, the result of % of the self-consistency error (red-line) and the regular model (green-line) should be within this distribution. But as we can see, the % of errors of the self-consistency model is below the distribution of the percentage of errors in the mixed sample. While the percentage of errors in the regular model is above. The null hypothesis is rejected with p-value « 0.0033.

For most samples, in the experiment described above, if there were hallucinations in one model there would be in the other and vise-versa. The alignment in the presence of factual-errors between the model's output may relate to the base model performance itself or to the context of the document and biases in the datasets. In our manual evaluation we note several problems and biases in the XSum dataset (Narayan et al., 2018) that can cause these factual-errors. For example, articles that contain only the last name of some character contain hallucinations of the personal name of the character - as the personal name appears in the target summary.

4 Discussion - Dataset limitations

As was mentioned before, in a lot of documents only the last name of a character is presented, while in the target summary they add his first name. As we used a model, Pagsos (Zhang et al., 2019), fine-tuned on XSUM dataset (Narayan et al., 2018) - we also see that the generated summaries add a first name as the target does. But since this information is lacking in the source document, they are generating random first names. Furthermore, Sometimes the summaries contain external information that was not mentioned in the article. For example , if a name of a hospital appears in the document the summaries may contain the city or general location of the hospital although it did not appear in the original text. This may lead to external hallucinations.

In the XSUM dataset (Narayan et al., 2018), the summary was written alongside the document itself, by a professional, usually by the same author of the article. Therefore, he might have added/written small details inside the summary which was not mentioned explicitly in the article. This is important to note that this way of contracting the summaries affect the quality of the dataset and the summarization models trained on it, and expose it to such biases. For some sub-topics, such as sports articles, the prior knowledge of the author of the summary seems more significant than others.

5 Conclusion

Self-consistency can be used as a decoding strategy to mitigate hallucination and factual errors. Although the intuition behind paraphrasing is clear and simple, it did not outperform nor complement above the self-consistency improvement, as far as our current research work can tell. We also encourage the research community to develop better datasets for abstractive summarization and to suggest a fine-grained version of the XSUM one.

Limitations

In our evaluation process, we manually classified types of hallucination/factual errors according to Pagnoni et al. (2021) topology of factual errors.

This classification task was hard to do for non-expert. There was not always a clear cat. It should be made by more experienced linguists or by crowd-sourcing and not by a single student. We also only tag a small portion of the test-set. Which was enough for general analysis like we did. Although we wished for a more fine-grained analysis, that we can not do using this limited test set. For example, if we had a good way of classifying hallucination, we would further investigate which prompts, words or versions of the input article trigger hallucination. A distinction one can do when only varying the input while saving the content and the model itself.

The gain from inference-methods such as self-consistency can be larger as the model's scale grows (Wang et al., 2023). We also note from manual evaluation, and from prior knowledge that the type of hallucinations may change dramatically with model scale. So, the result of our evaluation and the relative improvement of the mitigation method may vary (we assume for the better) when scale increases. But we were limited in school resources in this aspect. The quality and diversity of the paraphrasing model can impact significantly as well, and also require more resources.

Acknowledgements

Thanks to Ben Bogin, for exposing us to the field, with interesting articles and ideas in this work-line, and for the fruitful discussion. Thanks to the writers of "Self consistency for open-ended generation" Jain et al., 2023, for sharing with us technical details following their article.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2017. Quantization and training of neural networks for efficient integer-arithmetic-only inference.

Siddhartha Jain, Xiaofei Ma, Anoop Deoras, and Bing Xiang. 2023. Self-consistency for open-ended generations.

Ziwei Ji, 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.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization.

OpenAI. 2023. Gpt-4 technical report.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Maxim Kuznetsov Vladimir Vorobev. 2023. A paraphrasing model based on chatgpt paraphrases.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

A Appendix

Semantic Frame Errors	PredE	The predicate in the summary statement is inconsistent with the source article.
	EntE	The primary arguments (or their attributes) of the predicate are wrong.
	CircE	The additional information (like location or time) specifying the circumstance. around a predicate is wrong.
Discourse Errors	CorefE	A pronoun/reference with wrong or non-existing antecedent.
	LinkE	Error in how multiple statements are linked together in the discourse (for example temporal ordering/causal link).
Content verifiability errors	OutE	The statement contains information not present in the source article.
	GramE	The grammar of the sentence is so wrong that it becomes meaningless.
OthE		Factual error that does not fall into the other categorizes.
Ne		No error

Table 3: Topology of factual errors as defined in Pagnoni et al. (2021)

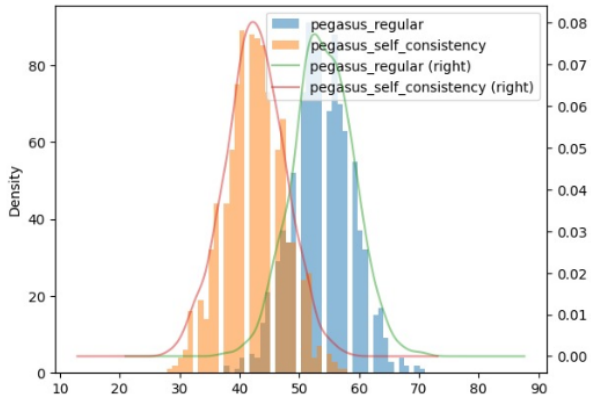

Figure 4: **Mitigating hallucinations - does self consistency improve the % of hallucination in model output?** To answer this question, we took bootstrap samples from each model output and calculating the mean % of errors in the sampled test-set (from the total 1% of test set we tagged). In this plot, we can see the histogram (left) and pdf (right) of the distribution of this sampling. Note, there is quite a large overlap between the distributions. Which means the results are not statistically significant. But this bootstrapping method does not take into account the fact that model outputs are dependent (if there was error in one it's likely to have error in the other). Also, because we calculate the mean of the samples with replacement - the samples are correlated and not i.i.d, so we can not use t-test or alike to evaluate statistical-significant results in this case. We can only see that self-consistency might improve the % of errors in its test sample, compared to the original model.