Article

Metric Ensembles for Hallucination Detection

Grant C. Forbes[1], Parth Katlana [2] and Zeydy Ortiz [3]

[1] North Carolina State University; gforbes@ncsu.edu
[2] North Carolina State University; pkatlan@ncsu.edu

Abstract: Abstractive text summarization has garnered increased interest as of late, in part due to the proliferation of large language models (LLMs). One of the most pressing problems related to generation of abstractive summaries is the need to reduce "hallucinations," information that was not included in the document being summarized, and which may be wholly incorrect. Due to this need, a wide array of metrics estimating consistency with the text being summarized have been proposed. We examine in particular a suite of unsupervised metrics for summary consistency, and measure their correlations with each other and with human evaluation scores in the Wikibio GPT3 hallucination dataset. We then compare these evaluations to models made from a simple linear ensemble of these metrics. We find that LLM-based methods outperform other unsupervised metrics for hallucination detection. We also find that ensemble methods can improve these scores even further, provided that the metrics in the ensemble have sufficiently similar and uncorrelated error rates. Finally, we present an ensemble method for LLM-based evaluations that we show improves over this previous SOTA.

Keywords: Large Language Models; Text summarization; Hallucination Detection; Ensemble methods

Citation: Forbes, C. G.; Katlana, P.; Ortiz, Z. Title. *Journal Not Specified* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

2. Introduction

Text summarization is a rapidly changing and advancing field, due in no small part to the advent of Large Language Models (LLMs) such as GPT [1] and LaMDA [2]. Many summarization methods, however, struggle with "hallucinating," or inserting false, misleading and/or nonrepresentative material into the summaries. As such, many automatic methods for hallucination detection have been proposed in the literature for

3. Metrics Evaluated and Related Work

Here, we discuss prior work, and describe the metrics we've chosen to represent the suite of prior methods that exist. We also discuss ensemble methods, and the conditions generally required for them to be effective.

3.1. *Text Summarization*

Traditionally, text summarization has been broken into two main methods: extractive and abstractive. Extractive text summarizers, such as OCCAMS [3], produce summaries by pulling particularly salient points from the document being summarized. On the other hand, abstractive summarizers, such as most attempts using large language models [4], attempt to generate a summary "from scratch," assembling new sentences in an attempt to synthesize the information in a document in a more human-seeming way. Abstractive summaries are often able to be more natural-sounding than extractive summaries, but, as has been noted repeatedly in the literature, have a tendency to hallucinate [5]. It is often challenging to evaluate these models, as has been noted in [6].

3.2. *Hallucination Detection Metrics*

There are many metrics for hallucination detection. Here are the ones we're using. Here, we are specifically concerned with unsupervised hallucination detection methods:

that is to say, those which require no input other than the summary and the source text itself. These metrics are the most general, requiring no gold-standard human summaries or other supplementary information, and are thus the most widely deployable. More particularly, unsupervised hallucination detection metrics are deployable in two important context which exclude any other types of metrics:

1. As an evaluative tool on summarization data (possibly generated continuously, rather than part of a finite set)

2. As an in-the-loop tool for actively curbing hallucinations in a summarization tool at runtime.

We chose a suite of unsupervised metrics intended to cover the breadth of work in this area, as well as being wel-backed, both theoretically and empirically. The survey [7] identifies four general types of unsupervised summarization metrics, and we evaluate representatives from each of these categories. We chose these to both cover the breadth of evaluation methods in the literature (i.e., question-answer based vs natural language entailment based, etc.), as well as to find methods with good theoretical/empirical backing and wide use, while still being relatively timely/current/up to date. Cite and talk about them here. More discussion of factual inconsistency detection can be found in the survey here [7].

3.2.1. FactSum

Factsumm[8] is a model-based metric to estimate the factual accuracy of generated text that is complementary to typical scoring schemes like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy). It is pre-trained on a large-scale dataset based on Wikipedia and Wikidata to train relation classifiers and end-to-end fact extraction models. The end-to-end models are shown to be able to extract complete sets of facts from datasets with full pages of text. It also analyses Multiple models that estimate factual accuracy on a Wikipedia text summarization task, and show their efficacy compared to ROUGE and other model-free variants by conducting a human evaluation study. [9]

[10]

3.2.2. QAGS

QAGS [11] is a question-answering based metric that automatically generates questions and answers from the source document and summary, then scores the summary based on how many of the derived questions are answered correctly. We used the implementation of QAGS in [9]. Here is QAEval, for comparison: [12] and here is FEQA: [13].

3.2.3. Rouge

[14] While ROUGE is traditionally used as a supervised metric, by calculating the ROUGE score between a generated summary and a gold-standard human summary, the idea of using ROUGE as an unsupervised metric was introduced in [15], which also demonstrated its efficacy in such a capacity. ROUGE used in this unsupervised way involves taking the ROUGE score between the generated summary and the text itself being summarized.

3.2.4. SMART

SMART (Sentence MAtching for Rating Text) [16] evaluation works on two principal ideas. Firstly, treat sentences as basic units of matching between system and reference summaries, instead of tokens. At sentence-level, exactly matching sentences are most likely nonexistent (in datasets with abstractive reference summaries), thus SMART utilizes soft-matching functions to compare sentences. Secondly, SMART allows to compare the candidate system summary with both the reference summary and the source document. This is particularly important when evaluating dimensions of summary quality that rely on the source document such as factuality.

3.2.5. SummaC

We benchmark two models of SummaC that utilize the NLI Pair Matrix to obtain a score for the generated summary. SUMMACZS (Zero-Shot): In SUMMACZS, the pair matrix is reduced to a one-dimensional vector by taking the maximum value of each column. This step retains the score for the document sentence that provides the strongest support for each summary sentence. The next step involves taking the mean of the produced vector, resulting in a scalar value used as the final model score for the entire summary.

SUMMACConv (Convolution): To address limitations of SUMMACZS, SUMMACConv reduces reliance on extrema values and takes the entire distribution of entailment scores for each summary sentence into account. The NLI Pair Matrix's columns are converted into fixed-size histograms, representing the distribution of scores for each summary sentence. The histogram is created by binning the NLI scores into evenly spaced bins. A learned convolutional layer processes each histogram to convert the distribution into a single score for the respective summary sentence.

[17]

3.2.6. SelfCheckGPT

SelfCheckGPT [18] is an unsupervised hallucination detection method that relies on the intuition that factual generated summaries are much more likely to be similar to each other than those which contain hallucinations, and that language models which are confident in their knowledge are likely to have much less diverse responses than those which are making things up. It utilizes a range of distance metrics to check generated summaries against each other for similarity, and shows that higher similarity to other generated summaries is highly correlated with human annotations for textual consistency.

3.2.7. LLM Self-Evaluation

[19]

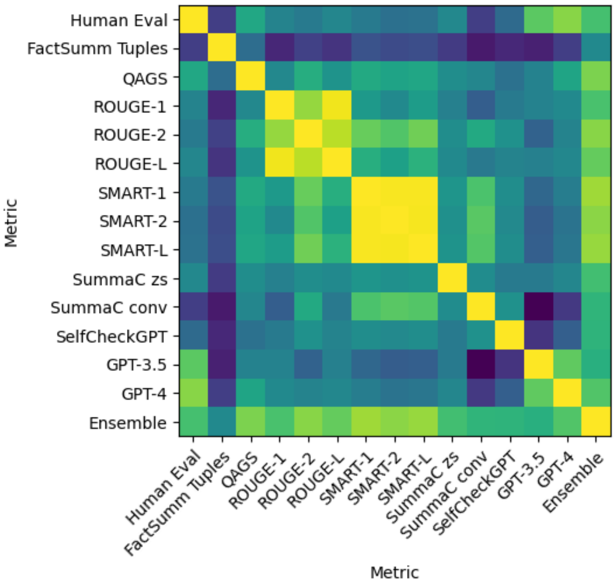

Figure 1

3.3. Metric Ensembles

It's been long noted that ensembles of models or metrics, even subpar ones combined naively, are surprisingly efficient and can rival or outperform expert judgement [20]. Ensembles also have been noted to aid in explainability in some contexts, something particularly relevant for analysis work, which is very sensitive to reliability and has a high threshold of required trust [21]. Particular in a domain such as hallucination detection, where there are a wide variety of disparate metrics, none of which is clearly superior to others in measuring the "true" value, ensemble methods are thus a promising avenue for improving over a baseline by leveraging the ability of each individual metric's error from the mean to cancel each other out. As derived in [22], if there is a collection of value-estimating functions f_i, each of which differs from some true function f by some $m_i = f - f_i$, and we assume the errors are uncorrelated, then in expectation we should expect the error m_{sum} of

$$f_{sum} = \frac{1}{N} \sum_{i=0}^{N-1} f_i \tag{1}$$

to be $\frac{\bar{m}_i}{N}$. So ensembles are a potentially powerful tool to deploy in spaces, such as hallucination detection where we have many diverse estimates for ground truth, but no (or prohibitively slow and expensive) access to that ground truth itself. There are then two qualities of some collection of metrics f_i that we would want, in order for an ensemble method to be effective:

1. The metrics must be diverse: that is, their errors should be relatively uncorrelated with each other.

2. The metrics must have m_i's that are similar in magnitude. If this condition is not met, then it is possible that the f_i with the lowest error alone would outperform the benefits of an ensemble model.

The metrics we're using in this work certainly meet condition 1., but it remains to be seen if they meet condition 2.

There has been some recent prior work in using ensemble-esque methods to leverage LLMs effectively. These often involve iterative prompting techniques, such as in [23], which prompts agents to "debate" each other before arriving at a final answer. SelfCheckGPT [18] itself could be seen as a variation of an ensemble method, as it involves self-checking the model's responses against other responses it might have given. Similar work is in [24], which incorporates a self-checking ensemble approach into the sampling algorithm for an LLM.

While in this work, we simply use naive similar ensembles of metrics with uniform weights, as has been shown to be effective [20,22], some other work has been done on using unlabelled data to find the best term weights for metrics [25–27]. While this particular line of work is applied specifically to binary classifiers, and we're working with metrics that cannot be generally constricted to outputs $\in \{0,1\}$, we believe something like this approach could be extended to metric ensembles in future work.

4. Method

We used the WikiBio GPT3 Hallucination dataset. This dataset consists of a subset of 238 entries from the original Wikibio dataset, generated in [28], accompanied by GPT3-generated summaries and sentence-by-sentence human evaluations of those summaries, ranking each sentence as "accurate," "minor inaccurate," or "major inaccurate." These additional summaries and human evaluations were generated in [18].

5. Results

Our results are split into roughly two sections: those comparing all metrics or ensembles of those metrics, and those which focus exclusively on "Non-LLM" metrics. Note that this term, as we're using it, refers to all metrics that do not involve a direct evaluation from an LLM: so SelfCheckGPT [18], while ostensibly running functions over many LLM outputs, is evaluated in only the former of these sections, and not the latter.

5.1. Non-LLM Metric Correlations

As a baseline, we first compute the mentioned unsupervised metrics and calculate their Pearson correlation with each other and the ensemble (table 1).

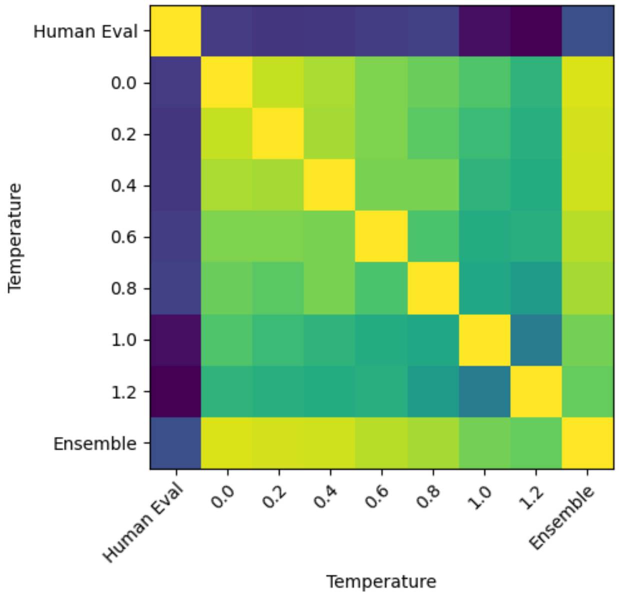

Figure 2

	Metrics			
	ROUGE-1	ROUGE-2	ROUGE-L	SummacCONV
SMART-1	0.72	0.86	0.77	0.85
SMART-2	1.00	0.50	0.75	0.80
SMART-l	1.00	0.50	0.75	0.90

Table 1. Pearson Correlation

Rouge and SMART Correlation: We observe a relatively high correlation between Rouge and SMART metrics. This strong correlation is attributed to their underlying similarity of measurement, which is based on overlaps between n-grams. This indicates that Rouge and SMART are capturing similar aspects of NLP evaluation and can be used interchangeably in certain cases.

article booktabs

SummacCONV and SMART Correlation: SummacCONV exhibits some correlation with SMART, although it is not as strong as the Rouge-SMART correlation. This suggests that SummacCONV shares some common ground with SMART in terms of evaluating NLP tasks but also has distinct characteristics that contribute to the moderate correlation.

Low Correlation of Other Metrics: On the other hand, many other metrics do not perform as well and demonstrate low correlation values. This implies that these metrics may measure different aspects of NLP evaluation compared to Rouge, SMART, and SummacCONV.

Given the low correlation of several metrics and the moderate correlation between SummacCONV and SMART, there is a clear indication for the need of an ensemble approach. An ensemble method can be utilized to combine the strengths of multiple metrics and improve the overall evaluation performance for NLP tasks. This will help in obtaining a more comprehensive and robust assessment of the models or systems under evaluation.

5.2. LLM Metrics Correlations

We compared the LLM across different temperatures for the prompt that stands to yield the most accurate response. [19]

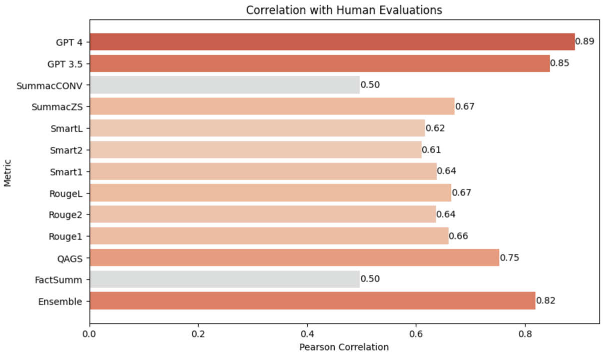

Figure 3

Metrics

Metrics		Human Eval	FactSumm Tuples	QAGS	ROUGE-1	ROUGE-2	ROUGE-L	SMART-1	SMART-2	SMART-L	SummaC$_{zs}$	SummaC$_{conv}$	SelfCheckGPT	GPT-3.5	GPT-4	Ensemble
Metrics	Human Eval	1.00	0.50	0.75	0.66	0.64	0.67	0.64	0.61	0.62	0.67	0.50	0.60	0.85	0.89	0.82
	FactSumm Tuples	0.50	1.00	0.60	0.45	0.50	0.47	0.54	0.52	0.53	0.49	0.43	0.46	0.44	0.50	0.68
	FactSumm Tuples	0.75	0.60	1.00	0.67	0.77	0.70	0.76	0.74	0.75	0.69	0.67	0.61	0.66	0.74	0.89
	ROUGE-1	0.66	0.45	0.67	1.00	0.90	0.99	0.72	0.68	0.72	0.65	0.57	0.63	0.66	0.68	0.82
	ROUGE-2	0.64	0.50	0.77	0.90	1.00	0.94	0.86	0.83	0.87	0.68	0.76	0.69	0.58	0.66	0.89
	ROUGE-L	0.67	0.47	0.70	0.99	0.94	1.00	0.77	0.73	0.78	0.67	0.64	0.66	0.65	0.67	0.85
	SMART-1	0.64	0.54	0.76	0.72	0.86	0.77	1.00	0.99	0.99	0.71	0.83	0.69	0.59	0.64	0.91
	SMART-2	0.61	0.52	0.74	0.68	0.83	0.73	0.99	1.00	0.99	0.70	0.84	0.68	0.56	0.62	0.89
	SMART-L	0.62	0.53	0.75	0.72	0.87	0.78	0.99	0.99	1.00	0.70	0.83	0.68	0.57	0.63	0.91
	SummaC$_{zs}$	0.68	0.49	0.69	0.65	0.68	0.67	0.71	0.69	0.70	1.00	0.68	0.64	0.64	0.67	0.81
	SummaC$_{conv}$	0.50	0.43	0.67	0.57	0.76	0.64	0.83	0.84	0.83	0.68	1.00	0.70	0.39	0.49	0.79
	SelfCheckGPT	0.60	0.46	0.61	0.63	0.69	0.66	0.69	0.67	0.68	0.64	0.70	1.00	0.48	0.57	0.78
	GPT-3.5	0.85	0.44	0.66	0.66	0.58	0.65	0.59	0.56	0.57	0.64	0.39	0.48	1.00	0.85	0.77
	GPT-4	0.89	0.50	0.74	0.68	0.66	0.67	0.64	0.62	0.63	0.67	0.49	0.57	0.85	1.00	0.83
	Ensemble	0.82	0.68	0.88	0.82	0.89	0.85	0.91	0.89	0.91	0.81	0.79	0.78	0.77	0.83	1.00

Table 2. Pearson Correlation

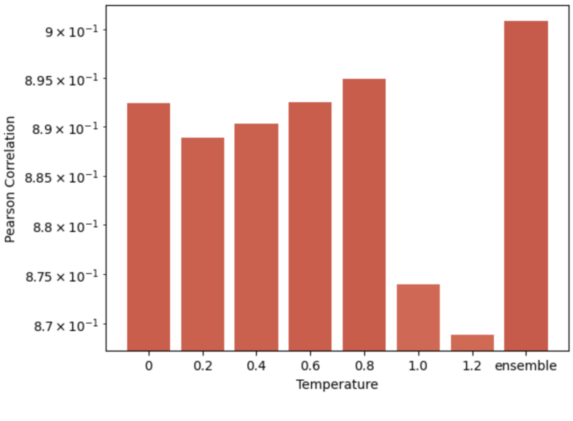

Figure 4

Previous literature has shown that LLMs perform better at lower temperatures and drastically decline in their efficiency at higher temperatures. We observe similar results in context to hallucination evaluations while using GPT 3.5turbox and GPT-4 for the ensemble in our study. The evaluation correlates increasingly from 0.2 to 0.8 before dropping drastically at temperature 1.0.

6. Discussion and Conclusion

Abstractive summaries are prone to hallucinations, meaning they may include statements that lack support from the original text. Some of these statements can be outright false, while others may be unsupported due to insufficient evidence within the source document. To address this issue, prior research has introduced several fact-checking tools that rely on automatic question-answering systems and textual entailment methods.[15,29]

In our study, we conducted a pilot experiment to explore the effectiveness of ensembles in detecting hallucinations. To evaluate their performance, we compared the ensembles using benchmark state-of-the-art metrics commonly employed in this domain. We have presented a simple self-training linear sum ensemble approach which leads to sizeable gains on both unsupervised metrics and LLMs evaluating unlabeled data for hallucination. We piloted the use of ensembles for hallucination detection by comparing them across the benchmark state-of-the-art metrics.

Improvements on the Benchmark. The models we introduced in this paper are just a first step towards harnessing ensemble models for hallucination detection. Future work could explore a number of improvements: measuring the errors for benchmarking with FRANK[30], optimizing weight redistribution to achieve the most optimal level, and creating a ground rule algorithm[27] by utilizing various metrics or combining multiple temperature settings.

Interpretability of model output. If a model has the ability to achieve better correlations with human evaluations or annotations, certain studies have indicated that ensemble models can proficiently quantify those problematic sections in many instances. Additionally, the ensemble can be further fine-tuned with respect to the temperatures along with other LLM models to establish consistency while scoring against other metrics.

Towards Consistent Summarization. Hallucination detection is but a first step in eliminating inconsistencies from summarization. Future work can include more powerful Hallucination detectors in the training of next-generation summarizers to both detect and reduce the prevalence of hallucinations in generated text

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: Please add: "This research received no external funding" or "This research was funded by NAME OF FUNDER grant number XXX." and and "The APC was funded by XXX". Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding, any errors may affect your future funding.

Institutional Review Board Statement: In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add "The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving humans. OR "The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving animals. OR

"Ethical review and approval were waived for this study due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans or animals.

Informed Consent Statement: Any research article describing a study involving humans should contain this statement. Please add "Informed consent was obtained from all subjects involved in the study." OR "Patient consent was waived due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state "Written informed consent has been obtained from the patient(s) to publish this paper" if applicable.

Data Availability Statement: We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: Declare conflicts of interest or state "The authors declare no conflict of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section. If there is no role, please state "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results".

Sample Availability: Samples of the compounds ... are available from the authors.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	Linear dichroism

Appendix A

Appendix A.1

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data are shown in the main text can be added here if brief, or as Supplementary Data. Mathematical proofs of results not central to the paper can be added as an appendix.

Appendix B

All appendix sections must be cited in the main text. In the appendices, Figures, Tables, etc. should be labeled, starting with "A"—e.g., Figure A1, Figure A2, etc.

References

1. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* **2023**.

2. Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* **2022**.
3. White, C.T.; Molino, N.P.; Yang, J.S.; Conroy, J.M. occams: A Text Summarization Package. *Analytics* **2023**, *2*, 546–559.
4. Zhang, H.; Liu, X.; Zhang, J. SummIt: Iterative Text Summarization via ChatGPT. *arXiv preprint arXiv:2305.14835* **2023**.
5. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys* **2023**, *55*, 1–38.
6. Goyal, T.; Li, J.J.; Durrett, G. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356* **2022**.
7. Huang, Y.; Feng, X.; Feng, X.; Qin, B. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839* **2021**.
8. Goodrich, B.; Rao, V.; Liu, P.J.; Saleh, M. Assessing the factual accuracy of generated text. In Proceedings of the proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 166–175.
9. Heo, H. FactSumm: Factual Consistency Scorer for Abstractive Summarization. https://github.com/Huffon/factsumm, 2021.
10. Kryściński, W.; McCann, B.; Xiong, C.; Socher, R. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840* **2019**.
11. Wang, A.; Cho, K.; Lewis, M. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228* **2020**.
12. Deutsch, D.; Bedrax-Weiss, T.; Roth, D. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics* **2021**, *9*, 774–789.
13. Durmus, E.; He, H.; Diab, M. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754* **2020**.
14. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text summarization branches out, 2004, pp. 74–81.
15. Louis, A.; Nenkova, A. Automatically evaluating content selection in summarization without human models **2009**.
16. Amplayo, R.K.; Liu, P.J.; Zhao, Y.; Narayan, S. SMART: sentences as basic units for text evaluation. *arXiv preprint arXiv:2208.01030* **2022**.
17. Laban, P.; Schnabel, T.; Bennett, P.N.; Hearst, M.A. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics* **2022**, *10*, 163–177.
18. Manakul, P.; Liusie, A.; Gales, M.J. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896* **2023**.
19. Luo, Z.; Xie, Q.; Ananiadou, S. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621* **2023**.
20. Dawes, R.M. The robust beauty of improper linear models in decision making. *American psychologist* **1979**, *34*, 571.
21. Forbes, G.; Crouser, R.J. Metric Ensembles Aid in Explainability: A Case Study with Wikipedia Data. *Analytics* **2023**, *2*, 315–327.
22. Perrone, M.P.; Cooper, L.N. When networks disagree: Ensemble methods for hybrid neural networks. In *How We Learn; How We Remember: Toward An Understanding Of Brain And Neural Systems: Selected Papers of Leon N Cooper*; World Scientific, 1995; pp. 342–358.
23. Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J.B.; Mordatch, I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv preprint arXiv:2305.14325* **2023**.
24. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* **2022**.
25. Platanios, E.A.; Blum, A.; Mitchell, T.M. Estimating Accuracy from Unlabeled Data. In Proceedings of the UAI, 2014, Vol. 14, p. 10.
26. Platanios, E.A.; Dubey, A.; Mitchell, T. Estimating accuracy from unlabeled data: A bayesian approach. In Proceedings of the International Conference on Machine Learning. PMLR, 2016, pp. 1416–1425.
27. Platanios, E.; Poon, H.; Mitchell, T.M.; Horvitz, E.J. Estimating accuracy from unlabeled data: A probabilistic logic approach. *Advances in neural information processing systems* **2017**, *30*.
28. Lebret, R.; Grangier, D.; Auli, M. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771* **2016**.
29. based self-training for abstractive opinion summarization, O.E. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. *arXiv preprint arXiv:2212.10791* **2022**.
30. Artidoro Pagnoni, Vidhisha Balachandran, Y.T. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. *arXiv preprint arXiv:2104.13346* **2021**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.