Laboratory for Analytic Sciences

Technical Report

Uncovering Factual Consistency Errors

Zeydy Ortiz [1], and Parth Katlana [2]

[1] DataCrunch Lab, LLC; zortiz@datacrunchlab.com
[2] North Carolina State University; pkatlan@ncsu.edu

Abstract: Automatic summarization using large language models has shown remarkable potential in generating concise and informative summaries from vast corpora of documents. However, a major concern is the introduction of factual consistency errors in these summaries, which can significantly impact the credibility and reliability of the information disseminated. In this study, we address the critical challenge of factual consistency checking, examining proposed methods and metrics to assess factuality. In particular, we utilize human evaluation to quantify the effectiveness of the proposed methods and evaluate how the metrics cover different types of factual consistency errors. The research can facilitate pre-deployment model selection and online verification of factual consistency, enhancing the trustworthiness of automatic summaries.

Keywords: factual consistency; hallucinations; summarization; large language models

1. Introduction

Delivering tailored daily reports (TLDRs) for individual knowledge workers will require producing high-quality summaries at scale. As natural language processing technologies continue to progress, large language models have shown remarkable proficiency in several tasks including summarization. However, a major concern with automatic summarization is the introduction of factual consistency errors. If the generated summaries contain statements that are not substantiated by the source documents, the credibility of TLDRs among knowledge workers will be compromised, ultimately undermining the entire effort. The inclusion of erroneous or inconsistent information can not only propagate misinformation but also erode the trust placed in automated systems. This, in turn, can hinder critical decision-making processes that heavily rely on the accuracy of the information provided. Therefore, upholding factual consistency stands as a paramount priority in ensuring the credibility and effectiveness of TLDRs.

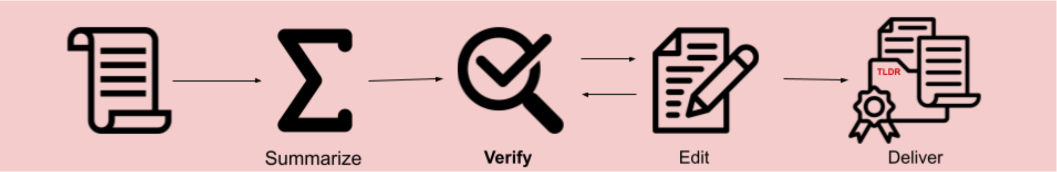

Figure 1. TLDR summarization pipeline.

During SCADS 2023, we contributed to **Critical Challenge 2.9** to "develop a technique to apply automatic fact-checking to abstractive summaries." We envision the summarization pipeline as depicted in Figure 1, where summaries undergo a thorough verification step before being incorporated in the TLDR. The summarization pipeline unfolds in the following manner: A document that was designated for the TLDR is processed by a summarization model, generating a summary. This summary is subsequently verified against the source document to detect any instances of factual inconsistency. If such errors are identified, the summary is revised and rechecked until all inconsistencies are rectified. Only once this verification process is complete, the refined summary is included in the TLDR

This technical report reflects work that was conducted as part of the 2023 Summer Conference on Applied Data Science (SCADS) hosted by the Laboratory for Analytic Sciences (LAS) at North Carolina State University.

for a knowledge worker. Since the goal is to produce summaries at scale, there is a clear imperative to automate the verification process for all the summaries.

To that end, we conducted analyses and empirical evaluations of various factuality metrics, including researching another area of interest for SCADS - "how might large language models like GPT-3 and GPT-4 be practically applied in the TLDR scenario". At present, the Holistic Evaluation of Language Models (HELM) [1] framework incorporates metrics for assessing the summarization task. This framework is currently used to rank summarization models based on their performance on a couple of metrics. We evaluated one of the metrics in HELM to help us address **Question 2.10** "how can we best adapt the tools from the HELM automatic evaluation to measure hallucination on our data at SCADS?" We then presented a set of recommended practices for effectively and automatically evaluating the factual consistency of summaries. These recommendations can be applied to **Critical Challenge 2.14** to "design/execute an experiment to identify and classify hallucination in automatically generated summaries." This research can be used to apply automatic checking of abstractive summaries in two areas: (a) pre-deployment testing to select a summarization model, and (b) online verification of summaries as part of the summarization pipeline. By automatically evaluating the factual consistency of summaries we can enhance the trustworthiness and usefulness of the TLDR, ultimately benefiting knowledge workers.

1.1. Terminology

We adopted the terminology introduced by Kryscinski et al [2] concerning the factual consistency of summaries, defining it as follows: *"a factually consistent summary contains only statements that are entailed by the source document."* Furthermore, they emphasize that *"factual consistency checking focuses on adherence of facts to information provided by a source document without guarantee that the information is true."*

According to these specific definitions, a summary that contains external world knowledge not present in the source document is considered **not** factually consistent, even if the facts from world knowledge are accurate. It is essential to make this distinction as our study does not encompass methods for fact-checking world knowledge.

Note that *hallucinations* are a type of factual consistency error. In particular, hallucinations are considered "out of article errors" by Pagnoni et al [3] and refer to information that can not be verified because is not present in the source document. This distinguishes hallucinations from other types of errors where there might be a discrepancy between the information discussed in the source document and the information presented in the summary. Further discussion on the typology of factual errors can be found in section 2.2 below.

2. Literature Review

The importance of factual consistency in automatic summarization has garnered considerable attention in the research community. Numerous studies have shed light on the challenges and limitations faced by language models when it comes to maintaining factual consistency in generated summaries[4–6].

Cao et al[7] estimate that about 30% of the summaries generated by sequence-to-sequence neural summarization models contain factual errors. Similarly, 30% of the summaries sampled in [8] contain factual consistency errors. Maynez et al [9] found that more than 70% of single-sentence summaries produced by a variety of neural network models contained factual errors. Goodrich et al [10] found that the summaries in their study of Wikipedia articles had factual inaccuracy rate of approximately 17%. Pagnoni et al [3] observed that 60% of the summaries in their study contain at least one factual error. Factual inconsistency varied between summarization models (23% to > 96% factually incorrect summaries) and datasets. They found that 43% of the summaries of CNN/Daily Mail contain errors while 92% of the summaries of the XSum dataset had errors.

2.1. Evaluating Factual Consistency

Factuality metrics serve as quantitative measures of the factual consistency of summaries. These metrics can be binary, classifying summaries as either factual or not factual, or they may provide a range of scores to account for various degrees of factual consistency. They score the summaries through a variety of techniques. Table 1 provides a summary of factuality metrics classified by the method they used to score summaries. We discuss these methods below. Pagnoni et al [3] benchmarked most of these metrics. Our study adds to their work by evaluating SMART, SummaC, and ChatGPT on their dataset.

Table 1. Types of Factuality Metrics.

Type	Metrics
Similarity-based	ROUGE, BLEU, METEOR, BERTscore, **SMART**
Fact-based	OpenIE, FactSumm
Question-Answering-based	FEQA, QAGS, QAFactEval
Entailment-based	DAE, FactCC, **SummaC**
LLM-based	**ChatGPT (gpt-3.5-turbo, gpt-4)**

2.1.1. Similarity-based Metrics

Traditional metrics to assess the quality of summaries, like ROUGE, BLEU and METEOR, have been used to try to assess factual consistency.

ROUGE measures the quality of summaries by computing overlapping lexical units - unigram (ROUGE-1), bigram (ROUGE-2), and longest common subsequence (ROUGE-L) - between the summary and source document. Similarly, BLEU (Bilingual Evaluation Understudy) measures the precision of n-grams up to a certain length (typically 4), rewarding longer n-gram matches more. METEOR (Metric for Evaluation of Translation with Explicit ORdering) incorporates additional linguistic and semantic features into the evaluation process. METEOR considers synonyms and stemming variations in addition to exact n-gram matches.

The intuition is that a summary that shares more lexical units with the source document, remains faithful to the source and does not introduce factual errors. Unfortunately, a summary can have high scores with these traditional metrics while still containing incorrect or misleading information. Another issue is that abstractive summaries that paraphrase the content are disadvantaged by these string-matching approaches.

To consider contextual information and capture semantic similarity, the BERTscore leverages embeddings from a pre-trained language model like BERT (Bidirectional Encoder Representations from Transformers). BERTscore measures the semantic similarity and contextual overlap between individual tokens (words or subwords) in the generated and reference texts. A high BERTscore suggests that the summary is semantically similar and contextually aligned with the source document.

Amplayo et al [11] introduced the idea of using sentences as basic units of matching instead of the traditional method of lexical units. They developed the SMART (Sentence MAtching for Rating Text) metric to assess four dimensions of summary quality: coherence, factuality, fluency, and informativeness. SMART has two types: n-gram overlap (SMART-n) and longest common subsequence (SMART-L), similar to the ROUGE metric. The score is calculated by comparing the candidate system summary with both the reference summary and the source document using a sentence-level soft-matching function. The authors studied various soft-matching functions for the SMART metric, string-based like CHRF and model-based leveraging embeddings. CHRF (Character n-gram F-score) is a machine translation evaluation metric that calculates character-based n-gram overlap between summary and source document sentences.

2.1.2. Fact-based Metrics

Fact-based methods to assess factual consistency extract facts from the source document and the generated summary utilizing named-entity recognition and relation extraction.

A fact is represented by a relation triple consisting of `(subject, relation, object)`. The extracted facts are then compared to determine factual consistency. OpenIE (Open Information Extraction) [12] extracts triples with an unspecified schema, and the relation is the text linking the two entities. Unspecified schema extraction makes the extracted triples hard to compare with each other because they need to be identical in order to match. In contrast, using a fixed schema [10,13] to predict the relations helps match the extracted facts.

2.1.3. Question-Answering-based Metrics

Another method to assess factual consistency is to automatically generate questions from a summary and answer them with the information in the source document. The Faithfulness Evaluation with Question Answering (FEQA) [14] method uses a QA model to predict answers from the source document. QAGS [15] compares the predicted answers from both the source document and the summary.

To determine the factual consistency of summaries, the Holistic Evaluation of Language Model (HELM) [1] framework includes QAFactEval [16]. QAFactEval was the focus of an initial study at SCADS 2022 and is not included in this study.

2.1.4. Entailment-based Metrics

Entailment-based metrics utilize natural language inference (NLI) to determine if a summary is entailed from the source document. Falke et al [4] experimented with out-of-the box NLI models and concluded that they do not perform well on the task. This is similar to our experience where the NLI model classified the majority of sentences as 'neutral.'

Goyal and Durrett [17] decompose entailment decision in a sentence in a more fine-grained way by using dependency arcs as semantic units and evaluating their entailment. Their method, Dependency Arc Entailment (DAE), had the highest correlation with discourse errors in [3].

Kryscinski et al [2] proposed a document-sentence approach for factual consistency checking, where each sentence of the summary is verified against the entire body of the source document. They used a weakly-supervised BERT-based model for verifying factual consistency, and added modules to identify the span of text in the source document and summary that correspond to the assessment. Their approach, named FactCC, performed best in the FRANK dataset that we used in this study.

Laban et al [18] re-examined the use of natural language inference (NLI) to assess factual consistency by segmenting documents into sentence units and aggregating scores between pairs of sentences. They introduced two factuality metrics, SummaC-zs and SummaC-conv, that start with an NLI Pair Matrix. The NLI Pair Matrix is produced by splitting a (document, summary) pair into sentence blocks. Each document and summary pair is run through the NLI model which produces a probability distribution over the three NLI categories (entailment, contradiction, and neutral). The resulting probabilities are used to populate the NLI Pair Matrix. SummaC-zs performs zero-shot aggregation by combining sentence-level scores into a single score for the entire summary. SummaC-conv is a trained model consisting of a single learned convolutional layer used to convert the entire distribution of entailment scores of all summary sentences into a single score.

They studied various models for the metrics including VITC, a combination of the Vitamin C[19] model and the MNLI (Multi-Genre Natural Language Inference) [20] model. The Vitamin C model is a pre-trained transformer-based model that was fine-tuned on a large-scale NLI dataset. The MNLI model is a pre-trained transformer-based model that was fine-tuned on the MNLI dataset, which consists of a diverse set of genres and domains.

SummaC is one of the metrics included in HELM to determine factual consistency.

2.1.5. LLM-based Metrics

An emerging method to evaluate factual consistency is to simply ask a large language model (LLM) to perform the evaluation. Luo et al [21] introduced ChatGPT as a tool to evaluate factual consistency through three distinct approaches: (1) entailment inference, (2) summary ranking, and (3) consistency ranking. Entailment inference consists of asking the LLM to determine if the summary is factually consistent with the source document and provide a binary response. Summary ranking consists of providing two summaries of the source document and asking the LLM to select the summary that is factually consistent. In consistency ranking, the LLM is asked to rate the summary given a range of values.

2.2. Typology of factual errors

Pagnoni et al [3] introduced a typology of factual consistency errors theoretically grounded in frame semantics and linguistic discourse analysis. In this typology, errors can arise at the semantic frame level, at the discourse level, or because the content can not be verified. Table 2 provides the definition of the various errors. We used these definitions for the ablation study.

Table 2. Typology of factual errors.

Level	Category	Description
Semantic Frame Errors	Entity Error (EntE)	Errors where the primary arguments (e.g. entities) of the predicate are wrong or have the wrong attributes
	Predicate Error (PredE)	Errors where the predicate in the summary is inconsistent with the source text
	Circumstance Error (CircE)	Errors where the circumstance around a predicate is wrong (e.g., location or time)
Discourse Errors	Coreference Error (CorefE)	Errors where pronouns and other references are incorrect or have no clear antecedent
	Discourse Link Error (LinkE)	Errors of incorrect temporal ordering or incorrect discourse links between statements
Content Verifiability Errors	Out of Article Error (OutE)	Errors where information can not be deduced from the original text (hallucinations)
	Grammatical Error (GramE)	Errors where the grammatical mistakes make the meaning of the statement incomprehensible or ambiguous

3. Methodology

To systematically evaluate factuality metrics, we utilized human evaluation as ground truth from the FRANK study [3]. We used the published results in the FRANK study for the factuality metrics they evaluated, namely, ROUGE, BLUE, METEOR, BERTScore, FEQA, DAE, QAGS, and FactCC. We evaluated sentences against the source document using the methods for SMART, SummaC, and ChatGPT metrics. In this section, we discuss the dataset, the factuality metrics under consideration, and the performance evaluation. Code, datasets and results available internally at https://github.ncsu.edu/SCADS/Factual_Consistency

3.1. Dataset

In this study, we used the FRANK dataset created by Pagnoni et al [3]. The dataset includes 250 news articles each from the CNN/Daily Mail [22] and XSum [23] datasets. They include publicly available summaries from five abstractive summarization models for the CNN/Daily Mail articles and from four models for the XSum articles. We selected this dataset since it has annotations from three human evaluators indicating the type of error found from the typology of errors in Table 2. They report an inter-annotator agreement Fleiss Kappa $\kappa = 0.58$ on whether a sentence is factual or not, with p = 91% of annotators agreeing with the majority class, and $\kappa = 0.39$ when all three annotators agree that a sentence is not factual with p = 73.9% of annotators agreeing with the majority class.

3.2. Factuality Metrics

We conducted a sentence-by-sentence evaluation of each summary with SMART, SummaC, and ChatGPT (gpt-3.5-turbo, gpt-4). In addition, we used the published scores for ROUGE, BLUE, METEOR, BERTScore, FEQA, DAE, QAGS, and FactCC in https://

github.com/artidoro/frank for the evaluation. The factuality metrics under consideration were configured as follows:

- **SMART** - We compared each sentence of the summary to the source document when running SMART. We used CHRF as the sentence-level soft-matching function. We collected results running SMART for unigrams (SMART 1), bigrams (SMART 2), and longest common subsequence (SMART L).
- **SummaC** - We utilized the VITC model when running SummaC. We ran both the zero-shot and the convolution methods, denoted as SummaC-zs and SummaC-conv respectively in our results. For operator choice, we used `max` for operator 1 and `mean` for operator 2 in SummaC-zs.
- **ChatGPT** - We used the entailment inference approach to ask the LLMs to determine the factual consistency of sentence summaries. We utilized Open AI's commercial models - *gpt-3.5-turbo* and the newly released *gpt-4* - through their API with the prompt in Listing 1. For each independent call, the `summary` consisted of one sentence of the summary. The `context` was the corresponding source document. We also elicited explanations from the models with the prompt in Listing 2. We set the `temperature` parameter to 0 to get more deterministic results. Results from zero-shot prompt are denoted as gpt-3.5 and gpt-4 for the corresponding LLM; results from the prompt to elicit explanations are denoted as gpt-3.5x and gpt-4x, respectively.

Listing 1. Zero-shot prompt

```
You are a respected analyst and writer.  Carefully consider the following
summary.  Determine the factual consistency of the summary given the
context, focusing strictly on the facts presented.  Respond only with
’accurate’ if the summary is consistent with the context, ’not accurate’
otherwise.  Do not provide an explanation.
Summary:  {summary}
Context:  {context}
Consistent:
```

Listing 2. Prompt to elicit explanations

```
You are a respected analyst and writer.  Carefully consider the following
summary.  Determine the factual consistency of the summary given the
context, focusing strictly on the facts presented.  Respond with ’accurate’
if the summary is consistent with the context, ’not accurate’ otherwise,
and provide an explanation for your reasoning.  Format response in JSON
with keys ’evaluation’ and ’explanation’ only.
Summary:  {summary}
Context:  {context}
Consistent:
```

3.3. Performance Evaluation

The most reliable method to validate factuality metrics is to compare results to human evaluation as ground truth. We determined the performance of the metrics by first converting the results into a binary label - factual or not factual. Since the SMART and SummaC metrics provide a score that is not immediately interpretable, the threshold for binary classification needs to be determined to balance precision and recall. We analyzed the Receiver Operating Characteristic (ROC) Curve to determine the best threshold for each of the metrics independently. This represents the optimal performance of the metrics for the summaries evaluated. The Area Under the Curve (AUC) score provides insight into their performance.

We calculated the accuracy and balanced accuracy of the metrics to determine how well they do in the classification of factual consistency. For mission support, we look beyond accuracy metrics to better understand the metrics through the lens of the TLDR use case. We compared the metrics on precision and recall. Precision measures the ability of the metric to correctly identify factual sentences from all the instances it classifies as factual. A high precision value indicates that the metric has a low rate of false positives, which means that when it determines that a sentence is factual, it is likely to be correct. For the TLDR, we need a classifier with high precision such that the sentences classified as factual would not need any further processing.

Recall, also known as sensitivity or true positive rate, measures the ability of the metric to correctly identify all factual sentences among the total actual factual sentence (positive) instances. For the TLDR application, precision takes precedence over recall since factual sentences that are incorrectly labeled would undergo additional fact-checking. However, a classifier that marks all sentences as not factual is not desirable.

Following the error analysis in [3], we compare the correlation of the metrics with human evaluation. Since we are comparing multiple metrics, we use partial correlation analysis after applying the Hotelling-Williams test to determine confounders. The main goal is to find how well the metrics capture the different errors.

4. Results

We compare the published results for ROUGE, BLUE, METEOR, BERTScore, FEQA, DAE, QAGS, and FactCC to our results on the valid split of the FRANK dataset. We collected scores for the test split from gpt-3.5-turbo and gpt-4 (without explanations) but did not have time to run the other metrics. We first calibrate SMART and SummaC to determine the threshold for binary classification and then compare the metrics based on their balanced accuracy, precision, and recall. Finally, we determine correlations with human evaluation and partial correlations for the different types of errors found in the summaries.

4.1. Calibrating metrics with ROC-AUC

SMART and SummaC metrics provide a score that needs to be calibrated to determine the threshold to use for binary classification. Figure 2 shows the receiver operating characteristics curve of the SMART metrics. Note that SMART-1 and SMART-L follow a similar curve while SMART-2 is slightly better. This is reflected in their AUC score of 0.73 for the former metrics and 0.74 for the latter. Similarly, figure 3 shows the ROC curve for the SummaC metrics. In this case, we see a distinction between these metrics with an AUC of 0.77 for SummaC-zs and 0.82 for SummaC-conv. On the test set of the FRANK dataset, Laban et al [18] report ROC-AUC of 85.3 and 88.4 respectively.

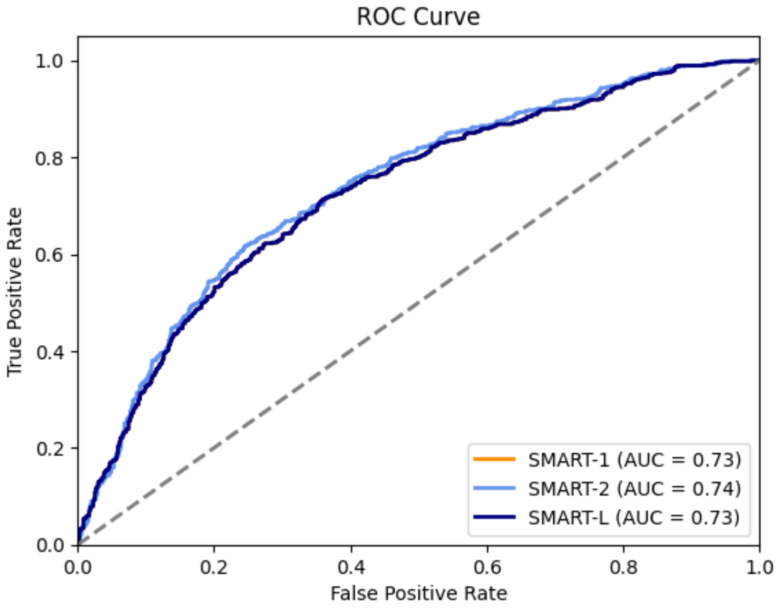

Figure 2. Receiver Operating Characteristics curve for SMART metrics.

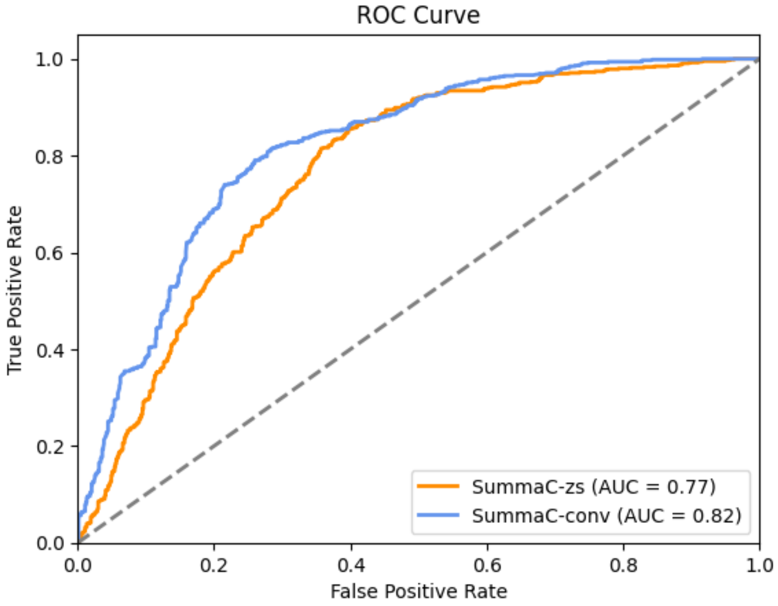

Figure 3. Receiver Operating Characteristics curve for SummaC metrics.

4.2. Performance Evaluation

Table 3 displays the performance results of FactCC (from the published scores), along with the results for SMART, SummaC, ChatGPT metrics. We computed the Pearson correlation between these metrics and human evaluations, as well as the accuracy, balanced accuracy, precision, and recall. A graphical representation of the balanced accuracy, precision, and recall results is presented in Figure 4. Notably, SummaC-conv exhibits the highest correlation with human evaluation. Consequently, it surpasses all other metrics in accuracy, balanced accuracy, and precision. It's worth mentioning that gpt-4 also demonstrates a strong correlation with human evaluation, resulting in performance akin to SummaC-conv. However, findings from the Wilcoxon Rank-Sum Test suggest that the groups are significantly different.

Table 3. Performance Results.

Metric	Pearson Correlation	Accuracy	Balanced Accuracy	Precision	Recall
FactCC	0.3882	0.63667	0.6691	0.5511	**0.9404**
SMART 1	0.3133	0.6279	0.6493	0.5514	0.8287
SMART 2	0.3431	0.6534	0.6688	0.5770	0.7966
SMART L	0.3133	0.6279	0.6493	0.5514	0.8287
SummaC-zs	0.4622	0.7133	0.7283	0.6284	0.8532
SummaC-conv	**0.5251**	**0.7584**	**0.7643**	**0.6918**	0.8135
gpt-3.5-turbo	0.1857	0.6057	0.5894	0.5646	0.4541
gpt-3.5-turbo-x	0.3255	0.6460	0.6605	0.5716	0.7813
gpt-4	0.4868	0.6756	0.7097	0.5762	**0.9939**
gpt-4-x	0.5115	0.6978	0.7287	0.5945	**0.9862**

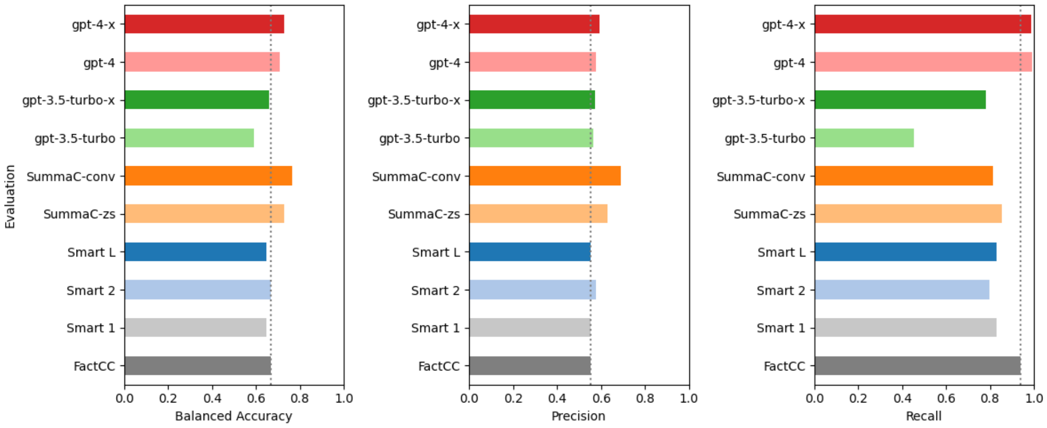

Figure 4. Performance of factuality metrics (**a**) Balanced Accuracy (**b**) Precision (**c**) Recall

For mission support, we look beyond accuracy metrics to gain a deeper understanding of the metrics in the context of the TLDR use case. While it is reassuring that most of the metrics have high recall rates, the precision of the metrics is a concern. We observe that these methods are not identifying all factual consistency errors in summary sentences; some of these summary sentences are mislabeled as 'factual'. We performed error analysis in 4.4 after evaluating metric-to-metric correlations and controlling for confounders below.

4.3. Correlations

Figure 5 shows a heatmap of the correlation between the different metrics considered in this study. As expected, the similarity-based metrics (ROUGE, BLEU, METEOR, and BERTScore) are correlated. Additionally, the SMART metrics have significant correlation between each other but not with the other similarity-based metrics. We noticed that gpt-4 and gpt-4x (with explanations) are correlated. They agree on their assessment 96% of the time. We believe that this is an indication of the reasoning power of gpt-4 - does not require additional prompting to elicit chain-of-thought reasoning.

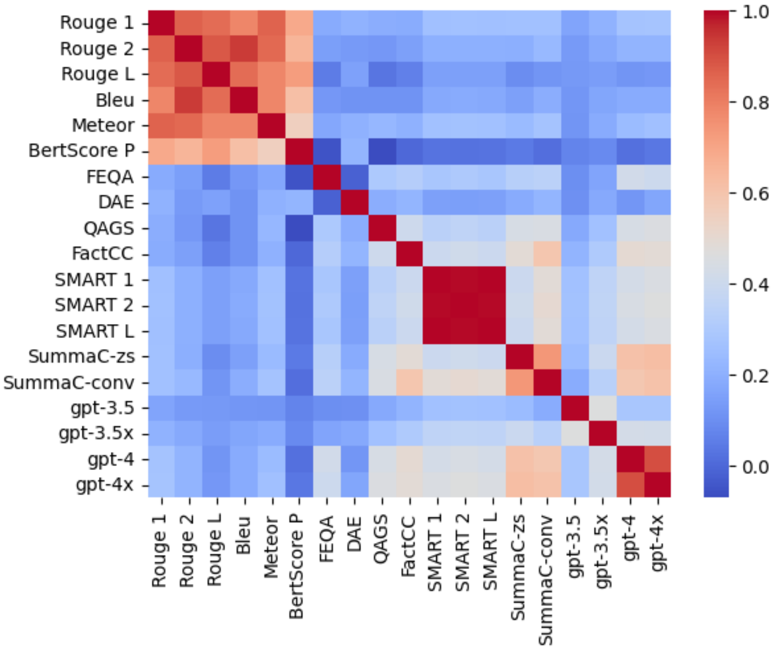

Figure 5. Metric-to-metric correlation of factuality metrics.

To effectively compare metrics and control for confounders, we performed the Hotelling-Williams test to find out which metrics are the same significant to the 0.05 threshold. Table 4 shows the results of the test.

Table 4. Metrics considered the same based on the Hotelling-Williams test.

Metric	Correlated Metrics
ROUGE 1	METEOR, FEQA
ROUGE 2	FEQA, DAE, gpt-3.5
ROUGE L	DAE, gpt-3.5
BLEU	DAE, gpt-3.5
METEOR	ROUGE 1, FEQA
BERTScore P	*None*
FEQA	ROUGE 1, ROUGE 2, METEOR
DAE	ROUGE 2, ROUGE L, BLEU, gpt-3.5
QAGS	gpt-3.5x
FactCC	SMART 1, SMART 2, SMART L
SMART 1	FactCC, SMART L
SMART 2	FactCC
SMART L	FactCC, SMART 1
SummaC-zs	gpt-4
SummaC-conv	*None*
gpt-3.5	ROUGE 2, ROUGE L, BLEU, DAE
gpt-3.5x	QAGS
gpt-4	SummaC-zs
gpt-4x	*None*

Based on that test, we then computed the partial correlation controlling for confounders. Figure 6 shows the partial correlation of the metrics of interest and human evaluation. We included FactCC as a reference metric since it was the best performing metric from the FRANK study. Notice that, after controlling for confounders, SMART 1 and SMART L had 0 partial correlation with human evaluation. As noted in Table 4, all SMART metrics are significantly correlated with FactCC. In addition, SMART 1 and SMART L are correlated with each other.

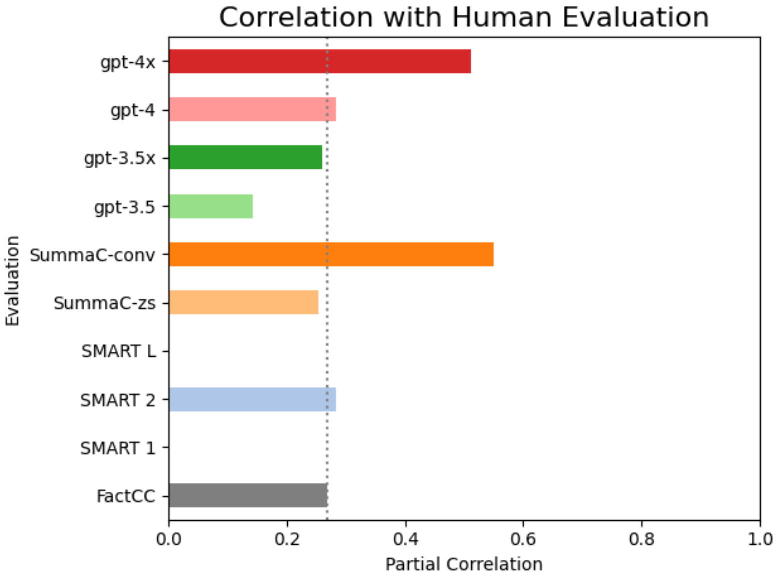

Figure 6. Partial correlation between factuality metrics and human evaluation.

4.4. Error Analysis

Guided by the research in the FRANK study [3], we conducted an evaluation of the partial correlation between the metrics of interest and the various types of errors identified in Table 2. Figures 7, 8, and 9 illustrate that relationship for the semantic frame, discourse, and content verifiability errors, respectively. As a baseline, we compare the results to FactCC.

Consistent with the performance results discussed earlier, notice that SummaC-conv exhibits the highest partial correlation across all error types, closely followed by gpt-4x. These metrics are significantly better than FactCC for all types of errors. Interestingly, these metrics demonstrate a strong correlation with content verifiability errors, particularly in the case of out of article errors (hallucinations). They are also better correlated to discourse errors than the metrics in the FRANK study.

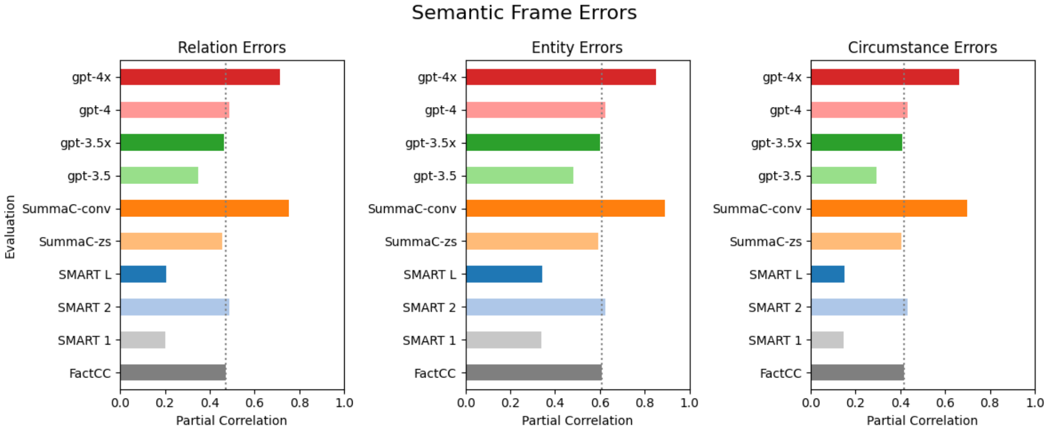

Figure 7. Partial correlation between factuality metrics and semantic frame errors (**a**) relation errors, (**b**) entity errors, and (**c**) circumnstance errors.

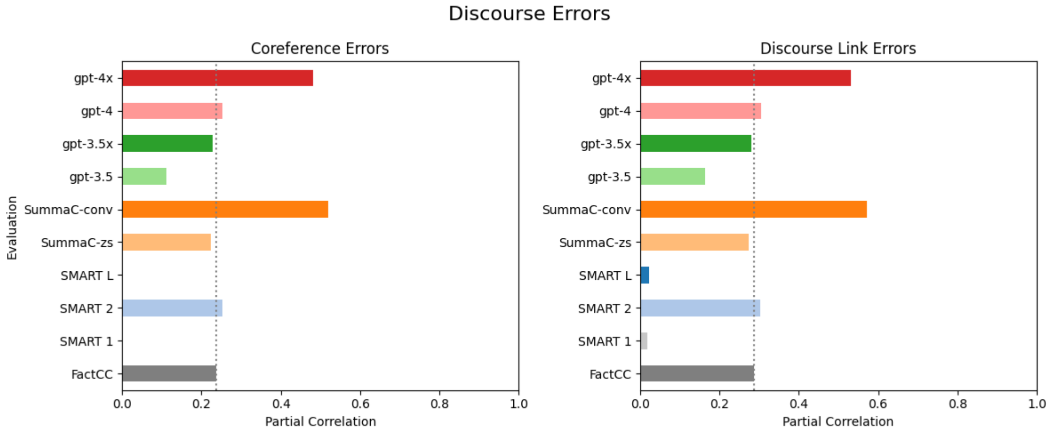

Figure 8. Partial correlation between factuality metrics and discourse errors (**a**) coreference errors, and (**b**) discourse link errors.

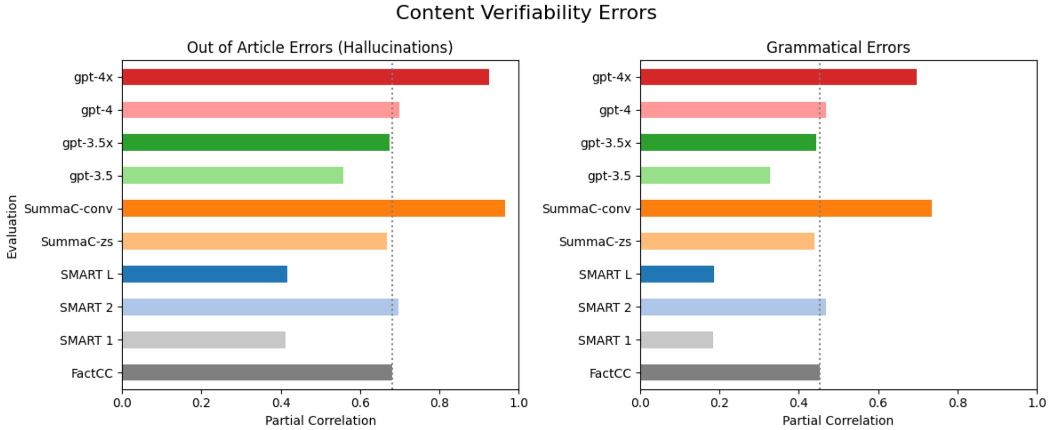

Figure 9. Partial correlation between factuality metrics and content verifiability errors (**a**) out of article errors (hallucinations), and (**b**) grammatical errors.

Analyzing the results presented, we observe that SummaC-conv demonstrates the most effective performance in factual consistency evaluation, closely followed by gpt-4x. However, it's important to note that the metrics, despite their strengths, do not capture all types of errors comprehensively. Particularly challenging are the identification of discourse errors, which prove to be elusive. Additionally, we note that SummaC-conv and gpt-4x exhibit minimal correlation with other factuality metrics, indicating unique strengths. This opens the possibility of combining these metrics with others to collectively enhance accuracy and precision.

5. Recommendations

5.1. Best practices in testing for factuality

Based on our experience validating metrics for factual consistency and our understanding of the TLDR, we recommend a number of best practices when testing for the factual consistency of summaries:

- **More than hallucinations** - Factual consistency of summaries can be influenced by various types of errors - from semantic frame errors and discourse errors to content verifiability errors. It's worth emphasizing that even when creating extractive summaries, factual consistency errors can arise if the construction process is not executed properly.
- **Sentence as a unit for evaluation** - We recommend dividing summaries into sentences and evaluating them against the source document. This approach has two benefits: (1)

provide adequate context for evaluation, and (2) localize the errors. Earlier efforts of comparing summary and source documents at the lexical units (unigrams, bigrams, etc.) would not appropriately evaluate summaries that were semantically similar. On the other end of the spectrum, evaluating entire summaries as a unit, would dilute the score and make it difficult to determine where errors might be in the summary. Some methods may also require reducing the source document to a few sentences in order be more precise in the evaluation. This may be necessary as the length of the document increases from the short documents in this study.

- **Human evaluation** - Utilize metrics that have been extensively validated against human evaluations as ground truth. We recommend evaluating summaries for documents that are similar to the type of document that would be included in TLDR.
- **Calibration of metrics** - Many factuality metrics are not directly interpretable. They provide a numeric score that needs to be converted into a binary label through the careful selection of a threshold value. Again, it is important to use human evaluation for the calibration on a variety of documents that reflect the use case. While we selected a threshold that maximizes balanced accuracy, a threshold that favors precision over recall may be desirable.
- **Representative documents** - Factual consistency errors varied across datasets and summarization models. Although using open-source data for benchmarking provides valuable insights, we recommend conducting tests with documents that closely resemble those intended for the TLDR. This includes considering factors like vocabulary, language, length, and other relevant characteristics.

5.2. Application to the TLDR

During SCADS 2023, we made progress in a number of challenges related to the TLDR. We believe the results we obtained can be readily integrated into the forthcoming TLDR system. Presented below are actionable recommendations derived from our findings that pertain to the identified challenges.

- **Critical Challenge 2.9 - Develop a technique to apply automatic fact-checking to abstractive summaries.**
  This research can be used to apply automatic checking of abstractive summaries in two areas: (a) pre-deployment testing, and (b) online verification of summaries as part of the summarization pipeline. During pre-deployment, the factuality metrics can be used to select the summarization model that is best at preserving the factual consistency of source documents. It is important to remember that these factuality metrics are designed to check consistency with source document but not to check world knowledge to verify that the facts are true. While a summarization model may reduce factual consistency errors in pre-deployment testing, we believe that their output should be checked during deployment. When the TLDR system is built, the factuality metrics can be used to verify the summaries before inclusion in the TLDR as envisioned in the summarization pipeline illustrated in Figure 1.
- **How might large language models like GPT-3 and GPT-4 be practically applied in the TLDR scenario.**
  We demonstrated the feasibility of using LLMs to uncover factual consistency errors in summaries. Their performance is competitive with the best metric. One of the benefits of using LLMs is that the assessment is immediately interpretable since they could respond with ’factual’ or ’not factual.’ Furthermore, eliciting explanations may provide insights on the errors found and how to correct them. However, if LLMs will be used to evaluate factual consistency, we recommend further testing and evaluation to determine their reliability.
- **Question 2.10 - How can we best adapt the tools from the HELM automatic evaluation to measure hallucination on our data at SCADS?**
  HELM could be used in pre-deployment testing to rank summarization models. Similar to our effort, HELM uses source documents from the CNN/Daily Mail and the

XSUM datasets for the summarization task. They evaluate the faithfulness of generated summaries using SummaC and QAFactEval. Their framework is designed to guide the selection of summarization models based on how they perform on these metrics.
However, for effective use in the TLDR system, additional metrics and representative datasets may need to be added to HELM for pre-deployment testing. In our study we observed that SummaC, in particular SummaC-conv, outperformed all other factuality metrics in the various performance assessments. However, to uncover errors that are missed by SummaC-conv, we recommend incorporating additional metrics to increase precision. In addition, the limitation of using news articles for summarization may need to be addressed in order to reflect the use case.

- **Critical Challenge 2.14 - Design/execute an experiment to identify and classify hallucination in automatically generated summaries.**
  In addition to the FRANK dataset, multiple open-source datasets have been annotated to assess factual consistency. We recommend leveraging these datasets to broaden the testing of factuality metrics and to rank summarization models before selection.
  Once the factuality metrics have been chosen, we recommend constructing a dataset that contains representative documents and summaries from the top-ranked summarization models. This dataset can then be annotated by human evaluators through the careful design of an experiment to identify factual consistency errors. This approach optimally utilizes the existing open-source datasets and maximizes the value of human evaluation.

5.3. Future Research

To continue making progress toward the TLDR, we identified several research opportunities. They include supporting the recommendations above as well as other research directions related to the summarization pipeline.

- **Extend research to select factuality metrics** - To make progress toward the TLDR, we recommend extending the research to select a metric or metrics to cover the types of factual errors that may arise in summaries. A combination of metrics might yield more effective results. Extending the research consists of obtaining scores from the test split of FRANK dataset, evaluating the metrics across diverse datasets, constructing a representative dataset with human evaluation to assess the performance of the metrics, and finally, combining different metrics to comprehensively address all error types.
- **Use factuality metrics to select summarization models** - Large language models continue to rapidly evolve. While we focused on determining how to evaluate summaries, future research may focus on reducing factual inconsistency in the summarization task through better training of the models or through fine-tuning pre-trained models. The selected factuality metrics can then be used during pre-deployment testing to rank summarization models.
- **Use factuality metrics to edit summaries** - The ultimate objective is to generate summaries that are factually consistent with the source documents. Factuality metrics, particularly those based on LLMs, offer valuable insights into enhancing the consistency of summaries. By assessing factuality on a sentence-by-sentence basis, it becomes simple to identify where errors arise within a summary. We collected explanations from gpt-3.5-turbo and gpt-4 for the valid split. Future research could explore leveraging these explanations to refine and edit the generated summaries.

6. Conclusions

Automatically evaluating the factual consistency of summaries offers a means to enhance the credibility and value of TLDRs. In this study, we explored existing approaches for assessing the factual consistency of summaries and provided an insightful evaluation of factuality metrics within the context of automatic summarization. The utility of factuality metrics is twofold within the proposed summarization pipeline. Primarily, these metrics play a pivotal role in pre-deployment testing, aiding in the selection of summarization models that mitigate factual consistency errors in generated summaries. Furthermore, these metrics are designed to be used during the actual summary generation process, ensuring accurate TLDR content. We contributed to a number of critical challenges in delivering TLDRs and provided recommendations for factuality testing, producing the TLDR, and future research.

Abbreviations

The following abbreviations are used in this manuscript:

BLEU	Bilingual Evaluation Understudy
CHRF	Character n-gram F-score
FEQA	Faithfulness Evaluation with Question Answering
HELM	Holistic Evaluation of Language Model
LAS	Laboratory of Analytic Sciences
LLM	Large Language Model
MNLI	Multi-Genre Natural Language Inference
NLI	Natural Language Inference
OpenIE	Open Information Extraction
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SCADS	Summer Conference on Applied Data Science
SMART	Sentence MAtching for Rating Text
SummaC-zs	SummaC - zero shot
SummaC-conv	SummaC - convolution
TLDR	Tailored Daily Report
VITC	Vitamin C and MNLI models

References

1. Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. Holistic Evaluation of Language Models, 2022, [arXiv:cs.CL/2211.09110].
2. Kryscinski, W.; McCann, B.; Xiong, C.; Socher, R. Evaluating the Factual Consistency of Abstractive Text Summarization. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Association for Computational Linguistics: Online, 2020; pp. 9332–9346. https://doi.org/10.18653/v1/2020.emnlp-main.750.
3. Pagnoni, A.; Balachandran, V.; Tsvetkov, Y. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Association for Computational Linguistics: Online, 2021; pp. 4812–4829. https://doi.org/10.18653/v1/2021.naacl-main.383.
4. Falke, T.; Ribeiro, L.F.R.; Utama, P.A.; Dagan, I.; Gurevych, I. Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Florence, Italy, 2019; pp. 2214–2220. https://doi.org/10.18653/v1/P19-1213.
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 Comput. Surv.* **2023**, *55*. https://doi.org/10.1145/3571730.
6. Huang, Y.; Feng, X.; Feng, X.; Qin, B. The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey, 2023, [arXiv:cs.CL/2104.14839].
7. Cao, Z.; Wei, F.; Li, W.; Li, S. Faithful to the Original: Fact Aware Neural Abstractive Summarization, 2017, [arXiv:cs.IR/1711.04434].
8. Kryscinski, W.; Keskar, N.S.; McCann, B.; Xiong, C.; Socher, R. Neural Text Summarization: A Critical Evaluation. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Association for Computational Linguistics: Hong Kong, China, 2019; pp. 540–551. https://doi.org/10.18653/v1/D19-1051.
9. Maynez, J.; Narayan, S.; Bohnet, B.; McDonald, R. On Faithfulness and Factuality in Abstractive Summarization. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Online, 2020; pp. 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173.
10. 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 &amp Data Mining. ACM, 2019. https://doi.org/10.1145/3292500.3330955.
11. Amplayo, R.K.; Liu, P.J.; Zhao, Y.; Narayan, S. SMART: Sentences as Basic Units for Text Evaluation, 2022, [arXiv:cs.CL/2208.01030].

12. Banko, M.; Cafarella, M.; Soderland, S.; Broadhead, M.; Etzioni, O. Open information extraction from the web in: Proceedings of the 20th international joint conference on artificial intelligence **2007**.
13. Heo, H. FactSumm: Factual Consistency Scorer for Abstractive Summarization. https://github.com/Huffon/factsumm, 2021.
14. Durmus, E.; He, H.; Diab, M. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.454.
15. Wang, A.; Cho, K.; Lewis, M. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. *ArXiv* **2020**, *abs/2004.04228*.
16. Fabbri, A.R.; Wu, C.S.; Liu, W.; Xiong, C. QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization, 2022, [arXiv:cs.CL/2112.08542].
17. Goyal, T.; Durrett, G. Evaluating Factuality in Generation with Dependency-level Entailment. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020; Association for Computational Linguistics: Online, 2020; pp. 3592–3603. https://doi.org/10.18653/v1/2020.findings-emnlp.322.
18. 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, [https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00453/1987014/tacl_a_00453.pdf]. https://doi.org/10.1162/tacl_a_00453.
19. Schuster, T.; Fisch, A.; Barzilay, R. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Association for Computational Linguistics: Online, 2021; pp. 624–643. https://doi.org/10.18653/v1/2021.naacl-main.52.
20. Williams, A.; Nangia, N.; Bowman, S. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Proceedings of the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); Association for Computational Linguistics: New Orleans, Louisiana, 2018; pp. 1112–1122. https://doi.org/10.18653/v1/N18-1101.
21. Luo, Z.; Xie, Q.; Ananiadou, S. ChatGPT as a Factual Inconsistency Evaluator for Text Summarization, 2023, [arXiv:cs.CL/2303.15621].
22. Hermann, K.M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; Blunsom, P. Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems* **2015**, *28*.
23. Narayan, S.; Cohen, S.B.; Lapata, M. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization, 2018, [arXiv:cs.CL/1808.08745].

Disclaimer/Publisher's Note: This material is based upon work supported in whole or in part with funding from the Laboratory for Analytic Sciences (LAS). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the LAS and/or any agency or entity of the United States Government.