QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization

Anonymous ACL submission

Abstract

Factual consistency is an essential quality of text summarization models in practical settings. Existing work in evaluating this dimension can be broadly categorized into two lines of research, entailment-based and question answering (QA)-based metrics, and different experimental setups often lead to contrasting conclusions as to which paradigm performs the best. In this work, we conduct an extensive comparison of entailment and QA-based metrics, demonstrating that carefully choosing the components of a QA-based metric, especially question generation and answerability classification, is critical to performance. Building on those insights, we propose an optimized metric, which we call QAFACTEVAL, that leads to a 14% average improvement over previous QA-based metrics on the SummaC factual consistency benchmark, and also outperforms the best-performing entailment-based metric. Moreover, we find that QA-based and entailment-based metrics can offer complementary signals and be combined into a single metric for a further performance boost.

1 Introduction

Text summarization aims to compress long document(s) into a short and fluent form that preserves salient information. The field has benefited from the application of pretrained methods (Liu and Lapata, 2019; Lewis et al., 2020; Zhang et al., 2020a). However, state-of-the-art models are not always factually consistent with the source documents they are conditioned on (Maynez et al., 2020; Fabbri et al., 2021). Thus, determining the factual consistency of a summary remains an essential task.

Recent metrics for summarization factual consistency can be broadly split into two categories: 1) Entailment-based metrics that determine whether the content in the summary is entailed by the input document (Kryscinski et al., 2020; Koto et al., 2020) and 2) QA-based metrics that compute a factual consistency score based on a QA model's ability to answer, using the input document, questions generated from the summary (Wang et al., 2020a; Durmus et al., 2020). We provide an illustrative example in Table 1 in which both metric types correctly identify the factual inconsistency and output a low score.

Document	
The Knicks beat the Rockets. The fans were excited.	
Summary	
The Knicks beat the Bucks.	
Entailment Matrix [Contra, Neutral, Support] $\begin{bmatrix} 0.90 & 0.07 & 0.03 \\ 0.02 & 0.90 & 0.08 \end{bmatrix}$	**Selected Answer** the Bucks **Generated Question** Who did the Knicks beat? **QA Output** the Rockets
Max Support Score 0.08	**Answer Overlap Score** 0.20

Table 1: Toy example of a factual inconsistency between a summary and a source document. *Left:* The entailment-based metric computes the level of contradiction, neutrality, and support between the summary and each source document sentence. The final factual consistency metric is calculated as the maximum support score over all source sentences. *Right:* The QA-based metric first selects a noun-phrase *answer* from the summary. A QG model then generates an associated question that a QA model answers based on the source document. The answer overlap score of the QA-based metric measures the semantic overlap between the QA model output and the selected answer as the final metric score.

Quantitative comparisons among entailment-based and QA-based metrics, however, often differ in their choices of baseline model and input granularity, evaluating on single datasets and drawing differing conclusions as to the best paradigm. For example, some work reports entailment-based metrics as performing best (Koto et al., 2020; Maynez et al., 2020), while other work argues for QA metrics (Durmus et al., 2020; Wang et al., 2020b; Scialom et al., 2021). Recently, Laban et al. (2021) proposed a benchmark called *SummaC* to compare metrics across six factual consistency datasets for the task of binary factual consistency classification, whether a summary is entirely factually consistent or not. This work unifies prior work on entailment-based metrics by studying the effect of input granularity, pretrained entailment model, and other hyperparameter choices on downstream evaluation performance. However, it does not study the components of QA-based metrics, which are more interpretable by their inherent decomposability.

To unify work in QA-based factual consistency evaluation, we do an extensive hyperparameter analysis of current metrics. We break down these metrics into four constituent components: 1) the selection of answers to ask questions about, 2) question generation (QG) conditioned upon these answers, 3) question answering (QA) based on the source document, and 4) answer overlap evaluation between QA model output and selected answers. We study the effect of each of these components on metric performance. Based on our insights, we propose an optimized metric, which we call QAFactEval , that outperforms the entailment-based metrics of Laban et al. (2021).

Our contributions are the following: 1) We analyze all components of the QA-based metric pipeline, and our proposed solution improves performance over prior QA-based metrics by over 14% on a factual consistency benchmark consisting of 6 individual datasets, achieving state-of-the-art results. 2) We show that QA-based metrics and NLI-based metrics offer complementary signals and combine them into a new metric via a simple learned network, further improving performance. 3) We report results for 10 additional metrics across classification and correlation analysis, providing the most comprehensive benchmark results for factual consistency metrics and highlighting areas for future work in QA-based metrics [1].

2 Related Work

Evaluating Factual Consistency Within entailment-based factual consistency evaluation, Falke et al. (2019) propose the task of ranking summary pairs for factual consistency based on entailment models, while Kryscinski et al. (2020) explore factual consistency classification jointly with source support or contradiction span extraction. Other work on entailment-based metrics has examined input granularity (Goyal and Durrett, 2020), trained on adversarial datasets (Barrantes et al., 2020), and explored entailment-based models as the backbone of others metrics such as BERTScore (Zhang et al., 2020b) as in Koto et al. (2021). Metric comparisons, however, were often conducted on isolated datasets. Laban et al. (2021) unify work in entailment-based metrics for factual consistency, showing the effect of granularity, base models, and other hyperparameter choices. This work also proposes a learned metric built on top of the output of an entailment model, with parameters fine-tuned on synthetic data. While this work fills a gap in the use of entailment-based metrics for factual consistency, our work analogously unifies QA-based metrics for factual consistency and proposes to combine entailment and QA-based metrics in a single learned metric.

QA-based evaluation metrics have received attention for summary quality dimensions beyond factual consistency (Eyal et al., 2019; Scialom et al., 2019; Deutsch et al., 2020). Recent work has shown that QA-based metrics better measure the overlap of information units for determining summary relevance over embedding-based metrics (Deutsch and Roth, 2021), further driving our study of QA-based metrics for factual consistency. While several QA-based metrics with similar structures have been applied for factual consistency, (Durmus et al., 2020; Wang et al., 2020b; Scialom et al., 2021), they differ in their underlying answer selection, question generation, question answering, and answer overlap components, reporting different performances. We perform a comprehensive evaluation of QA-based metric components and propose improved model components for the task of answer overlap and question filtering.

Summarization Benchmarking A recent line of work aims to take stock of the current state of summarization models and progress, both within factual consistency and across summarization more broadly. Kryscinski et al. (2019) note biases and failure modes of abstractive summarization models, while other work analyzes and collects annotations over the output of recent summarization models across multiple dimensions, including factual consistency (Fabbri et al., 2021; Bhandari et al., 2020; Huang et al., 2020). Lux et al. (2020) propose a typology of errors found in summarization models, while Gabriel et al. (2021) propose a framework for

[1] Code and metric outputs will be made publicly available: https://github.com/anonymous

meta-evaluation of factual consistency metrics. Laban et al. (2021) propose to combine recent work in factual consistency evaluation for summarization through a single benchmark. Our work directly makes use of this benchmark while emphasizing QA-based metrics. We also include correlation analysis for a more comprehensive understanding of current factual consistency metrics.

3 Evaluation Metrics

In this section, we introduce the factual consistency metrics studied, which we divide into entailment metrics, QA-based metrics, and learned metrics.

3.1 Entailment-based Metrics

The metrics below produce a score for each summary sentence which is then averaged to compute the final metric score.

MNLI applies a RoBERTa large (Liu et al., 2019) model trained on MNLI (Williams et al., 2018). The score of a summary sentence is the maximum entailment score over all input sentences.

ANLI Barrantes et al. (2020) uses the same method as the MNLI metric with a model trained on the ANLI (Nie et al., 2020) dataset consisting of adversarial datapoints.

SCZeroShot Laban et al. (2021) works analogously to the above metrics with a base model trained on both MNLI and Vitamin-C data (Schuster et al., 2021), consisting of closely-related contrastive entailment examples.

BertScore-FFCI Koto et al. (2021) applies BertScore (Zhang et al., 2020b) with an backbone RoBERTa-MNLI model, averaging the three highest BertScore F1 scores over the input sentences.

DAE Goyal and Durrett (2020) computes entailment scores between a source document and summary dependency arcs, applying an entailment model trained on synthetic data.

FactCC Kryscinski et al. (2020) is a RoBERTa-base model trained on FactCC synthetic data to compute a document-level score, and thus the scores need not be aggregated over input sentences.

DocNLI Yin et al. (2021) train a document-level entailment model, similar to the FactCC metric.

3.2 QA Metric Components

We now describe the components that constitute the QA-based pipeline for factual consistency. We refer to our metric, consisting of the best combination of the below components, as QAFactEval.

Answer Selection QA-based metrics compare information units between the summary and source, so it is thus necessary to first extract such units, or answers, from the given summary. We follow the protocols from Deutsch et al. (2020) and compare extracting the following answer types: named entities (*NER*), noun phrase chunks (*NP Chunks*), maximally sized noun phrases (*Max NP*), whereby the dependency subtrees of nouns reached by traversing a given sentence's dependency parse from the root are chosen as answers, and *All*, which combines answers from the above three techniques.

Question Generation Having selected answers, questions are generated conditioned upon these answers using the summary as context. Typically, this is an encoder-decoder model which inputs the answer and context separated by a special token. On the modeling side, we examine *BART* (Lewis et al., 2020) and *T5* (Raffel et al., 2019) as the underlying generators. On the data side, we experiment with models trained for question generation on *SQuAD* (Rajpurkar et al., 2016), a standard QA dataset consisting of questions on Wikipedia articles, and on *QA2D* (Demszky et al., 2018), a dataset of declarative sentences with associated question/answer pairs derived from SQuAD. Furthermore, we experiment with the recently-introduced *MixQG* models (Murakhovs'ka et al., 2021), which are T5 models trained on a combination of nine QA datasets with diverse answer types and which outperform other QG models across several tasks.

Question Answering The QA component answers questions from the previous steps using the input document as context. We experiment with both extractive QA models, which extract a text span from the input as an answer, and abstractive QA models, which generate an answer token-by-token. For extractive models, we ablate *Electra* (Clark et al., 2020), which achieves strong performance on the SQuAD 2.0 dataset, and *MADE* (Friedman et al., 2021), which models multi-dataset QA with a collection of dataset-specific adapter modules sharing the same underlying RoBERTa-base model. For abstractive QA, we experiment with *T5* fine-tuned on SQuAD and *UnifiedQA*

(Khashabi et al., 2020), an approach which trains a T5 QA model on 8 diverse, seed datasets and was shown to generalize across 20 datasets and 4 input formats. All QA models except MADE are trained on data containing unanswerable questions.

Answer Overlap Evaluation An answer overlap metric must be computed to determine the match between the initial answer selected in the first component and the QA model output. Typically, answer overlap in QA is measured through exact match (*EM*) score or word *F1* score. We also test a learned metric, the *LERC* score proposed by Chen et al. (2020). This metric outputs a 1-5 answer overlap score conditioned on a question and context. The scorer is trained on their MOCHA dataset, consisting of 40k crowdsourced judgments on QA model outputs. We include the BERT-base (Devlin et al., 2019) model from the original paper, which we call *LERC (orig)*. We additionally experiment with two models trained from RoBERTa-large checkpoints, one trained from the original checkpoint, *LERC (RoBERTa)*, and one initialized from Jia et al. (2021), which we call *LERC (QuIP)*, for the task of jointly encoding passages and answers with question-infused pretraining. Lastly, we experiment with the *IsAnsweredInput* answer metric, which is a 0/1 score of whether the question is answerable using the input document according to the QA model. We use the Electra-large QA model to determine whether a question is answerable, as this model shows strong performance on identifying unanswerable questions on SQuAD.

Question Filtering Model-generated questions may contain noise from the QG model itself or from disfluencies in the summary the QG model conditions upon. Such noisy questions can skew the overall metric score, as the QA component may be unable to correctly answer the question, regardless of the summary's factual consistency. We filter such questions through a step called *IsAnsweredSumm Filter*: the same Electra-large QA model returns a 0/1 score of whether the question is answerable, now using the summary as context, and questions labeled as unanswerable are filtered.

Overall For a given question, if IsAnsweredInput returns 0, the question is unanswerable using the input, we label all the above answer overlap scores as 0, and otherwise use the answer overlap score. We refer to this scoring of unanswerable questions as 0 as the *Answerability Penalty*. We also experiment with not setting the overlap score of these unanswerable questions to 0 but rather using the answer overlap score of the most probable answer from the QA model. Finally, the overall factual consistency score for each metric is computed as its average scores over all questions remaining following Question Filtering.

3.3 Learned Metrics

SCConv is a model introduced by Laban et al. (2021) which learns to aggregate entailment-model output scores across input sentences into a single score. More concretely, for a document consisting of M sentences and a summary consisting of N sentences, the entailment-based model produces an $M \times N$ matrix of entailment scores. The $M \times N$ matrix is then transformed to an $H \times N$ matrix by binning the M sentences to create a histogram, where H is the number of bins. This matrix is input to a 1-D convolution layer to produce a score for each summary sentence, and the scores are averaged across summary sentences. The parameters of this model are fine-tuned on synthetic data.

QAFACTEVAL-NLI While SCConv captures sentence-level support, QAFACTEVAL measures finer-grained answer overlap between the source and summary. Thus, we are able to combine these two into a single factual consistency metric, QAFACTEVAL-NLI. Assume that K answers are extracted from the summary. The pipeline described above will then output a single score per answer for the entire summary, resulting in an array of length K. We convert this to a histogram of size H in a similar manner as SCConv and pass this histogram through a 1-D convolution layer to produce a single QA score. This score is concatenated with the NLI score produced by SCConv and input to a linear layer to produce the final metric score. The linear layer can be trained in either *synthetic* or *supervised* ways, detailed in Section 4.2.

3.4 Additional Metrics

We include the following metrics for completeness.

BARTScore Yuan et al. (2021) calculates the log-likelihood from BART fine-tuned on CNN/DailyMail (Hermann et al., 2015; Nallapati et al., 2016) of the summary conditioned upon the source text as a metric for factual consistency.

BLANC Vasilyev et al. (2020) is a reference-less metric of summary quality that measures the difference in masked language modeling performance with and without access to the summary.

QuestEval (Scialom et al., 2021) is the prior state-of-the-art QA-based metric for factual consistency. The T5-base (SQuAD) QG and T5-base QA models described above are applied directly from the QuestEval metric. QuestEval generates questions based on the input document and answers them using the summary in addition to following the above QA metric pipeline. QuestEval aggregates the score from these two pipelines. We believe that our described pipeline more closely measures factual consistency, while generating questions from the source may confound factual consistency with relevance.

4 Methodology

We present the datasets explored for binary classification and correlation analyses. We also describe settings for reporting ablation and final results.

4.1 Data

The **SummaC** benchmark (Laban et al., 2021) introduces a collection of datasets for binary factual consistency evaluation. A data point is labeled as positive if it contains no factual inconsistencies or is rated the highest possible score in the case of Likert scaling, and as negative otherwise. We now briefly describe the datasets in the benchmark and any departures from the original benchmark, and additional datasets we use for correlation analysis. We refer the reader to Laban et al. (2021) for further details regarding the benchmark creation.

CGS Falke et al. (2019) consists of paired summary sentences from CNN/DailyMail (Hermann et al., 2015; Nallapati et al., 2016), one correct sentence and one containing an error. Laban et al. (2021) treats the correct summaries as positive examples and the others as negative examples.

XSF Maynez et al. (2020) consists of summaries from the XSum dataset (Narayan et al., 2018) annotated for word-level factual consistency errors.

Polytope Huang et al. (2020) propose a typology of eight summarization errors consisting of both content and stylistic errors and annotate model outputs from 10 systems on CNN/DailyMail data. The original SummaC benchmark included the Omission and Addition errors of this proposed typology as factual inconsistencies, but these are largely extractive, factually consistent summaries. We thus label these examples as factually consistent and report results on this modified dataset.

FactCC Kryscinski et al. (2020) introduce a factual consistency dataset on CNN/DailyMail annotated by the authors of the paper to ensure the quality of the annotations.

SummEval Fabbri et al. (2021) analyze summaries from 17 models on CNN/DailyMail across the dimensions of factual consistency, coherence, fluency, and relevance.

FRANK Pagnoni et al. (2021) introduce an extensive typology of errors made by summarization systems across CNN/DailyMail and XSum.

QAGs Wang et al. (2020b) crowdsource sentence-level summary annotations for factual consistency across CNN/Daily Mail and XSum data. We only report correlation analysis for this dataset as it was not a part of SummaC.

4.2 Experiment Setup

Metric Implementation Metrics were applied directly from the original GitHub repository or by using the SacreRouge Library (Deutsch and Roth, 2020), which was also used in correlation analysis. The learned metrics make use of code released from Laban et al. (2021) for training, and all models are implemented in PyTorch (Li et al., 2020) and in the Transformers library (Wolf et al., 2019). The BART-larege (QA2D) QG and Electra-large QA models are applied from the QAEval relevance modeling metric (Deutsch et al., 2020).

Ablation Settings Following Laban et al. (2021), a metric threshold score for binary classification is determined from the validation set of SummaC and applied to the test set. For ablation studies, we both perform thresholding and evaluation on the validation set to preserve the integrity of the test set. For each benchmark dataset, we sample a random subset of 80% of the validation set to determine the threshold and evaluate on the remaining 20% of the validation set. The best performing combination of QA metric components constitutes our QAFactEval metric. We take the best performing combination of QA metric components and vary a given component, such as answer selection, while holding all other components constant and consistent with the best component combination.

Training Settings To tune the parameters of the learned metrics, we train on a subset of 50k synthetic data points from FactCC, following Laban et al. (2021). We name these runs *synthetic* setting due to the lack of human-labeled data. We also experiment with a *supervised* setting by fine-tuning the parameters on the SummaC validation set for each individual dataset, choosing the threshold on this validation data, and applying the model to the test set. Training on such a small amount of data is feasible due to the small number of parameters of the learned metrics. Cross entropy loss with Adam (Kingma and Ba, 2015) optimizer is used, with a batch size of 32 and a learning rate of 1e-2.

5 Results

In this section, we first study the effects of model component choices on QAFactEval. We then compare metric results across both the SummaC binary classification task and correlation analysis.

5.1 Ablation Results

We provide the results of our ablation studies on the components of QA-based metrics in Table 2 and show two illustrative examples in Table 4.

Component	Model Choice	Benchmark
QAFactEval		**77.5**
Answer Selection	**NP Chunks**	-
	Max NP	75.7
	NER	66.4
	ALL	75.7
Question Generation	**BART-large (QA2D)**	-
	BART-large (SQuAD)	74.3
	T5-base (SQuAD)	67.0
	MixQG-base	75.1
	MixQG-large	74.9
Question Answering	**Electra-large**	-
	Electra-base	77.0
	MADE	77.4
	T5-base	76.1
	UnifiedQA-base	75.7
Answer Overlap	**LERC (QuIP)**	-
	EM	68.4
	F1	71.7
	IsAnsweredInput	73.3
	LERC (orig)	71.8
	LERC (RoBERTa)	77.3
Filtering/Answerability	**Both**	-
	No IsAnsweredSumm Filter	73.8
	No Answerability Penalty	72.1
	Neither	67.4

Table 2: Results of ablation studies on the SummaC benchmark validation set, showing the effect of the individual components of QAFactEval. The first row represents the performance of the best combination of components. Ablations are performed by swapping a given component while holding all others consistent with the best overall model, and the best setting is bolded.

Effect of Answer Selection Selecting NP Chunks performs best, aligning with Deutsch et al. (2020), which shows that NP Chunks obtain the largest coverage of information units while retaining high precision. We find a large decrease in performance when selecting NER and only a slight decrease in performance when choosing Max NP or ALL answers together. Named entity selection likely performs worse due to the scarcity of extracted answers; only three entities are extracted on average across the benchmark, while all other approaches extract over 10 answers per summary.

Effect of QG Models The choice of the QG model notably affects downstream performance. BART-large (QA2D) works the best and produces much longer questions, about 17 tokens on average, versus about 10 from the other models. Deutsch et al. (2020) note how humans tend to produce shorter questions. However, longer questions may be preferable for this task to facilitate the QA model's ability to understand and answer the question. BART-large (QA2D) also is the most extractive, with only about 20% novel unigrams in the question, while T5-base (SQuAD) model is the most abstractive with about 47% novel unigrams, resulting in occasional hallucinations and questions that the QA model struggles to answer. As seen in Table 4, MixQG models do often produce highly-fluent questions, but the longer, highly-extractive output of BART-large (QA2D) improves downstream factual consistency performance.

Effect of QA Model Surprisingly, we do not find a large difference in the QA model component across model sizes or between extractive and abstractive QA models, implying that QA ability is not the bottleneck of our task. In this setting, we keep IsAnsweredInput from Electra-large constant, as not all QA models are trained with unanswerable questions; thus the only differences are in the answers to questions marked as answerable.

Effect of Answer Overlap Metric We observe a large difference between EM and other overlap metrics. We also see a notable gap between LERC (orig) and LERC (RoBERTa) along with a further slight improvement with LERC (QuIP), showing the effect of the underlying model of the learned metric on factual consistency performance.

Effect of Question Filtering and Answerability Not filtering questions according to the QA model's ability to answer them using the summary decreases performance. Furthermore, not applying

Model Type	Model Name	CGS	XSF	Polytope	FactCC	SummEval	FRANK	Benchmark
Misc	BARTScore	63.3	53.3	80.4	66.8	69.8	80.0	68.9
	BLANC	51.6	54.5	72.2	53.0	63.0	76.2	61.8
Entailment	FactCC	64.8	56.6	80.2	77.1	73.6	70.3	70.4
	BertScore-FFCI	56.9	**68.8**	69.2	57.9	67.4	71.9	65.4
	DAE	71.3	49.7	78.9	80.7	74.7	81.0	72.7
	ANLI	74.9	53.0	77.6	85.8	75.9	78.9	74.4
	MNLI	67.6	61.5	77.3	89.8	78.7	79.6	75.7
	DocNLI	49.6	57.0	**84.7**	73.0	75.6	70.9	68.5
	SCZeroShot	59.6	56.1	81.5	83.2	77.9	78.5	72.8
QA	QuestEval	59.4	61.9	73.1	66.5	68.4	79.8	68.2
	QAFactEval	75.1	63.1	79.8	84.1	**80.9**	83.9	77.8
Learned	SCConv (synthetic)	60.8	60.9	76.0	88.1	78.1	81.6	74.3
	QAFactEval-NLI (synthetic)	74.2	59.1	82.1	**91.1**	80.2	83.4	78.3
	QAFactEval-NLI (supervised)	**78.1**	60.9	83.7	89.3	80.5	**84.3***	**79.5***

Table 3: Balanced accuracy on the test set of the six SummaC benchmark datasets, and the average over the benchmark. Metrics are divided into entailment-based, QA-based, and learned metrics that are fine-tuned on synthetic or supervised data. An improvement over prior work with a 99% confidence interval is indicated by *.

<table>
<tr><td>Document</td><td colspan="2">Paul Merson has restarted his row with Andros Townsend. ... '... it was a great goal,' Merson said. 'It's just a matter of opinion, and ... he got pulled off after half an hour .... in front of Roy Hodgson, so he shouldn't have been in the squad. ...' ... Sky Sports pundit Merson (centre) criticised Townsend's call-up to the England squad last week ....</td><td>They're not gonna take it anymore. Really. Twisted Sister says that its 2016 tour will be its last, according to a press release. ... The band will also perform two shows in Pero's honor: one at Las Vegas Hard Rock Hotel and Casino, the other at the Starland Ballroom in Sayreville, New Jersey.</td></tr>
<tr><td>Summary</td><td colspan="2">Paul Merson is not happy with Andros Townsend's call-up to the England squad last week</td><td>The band will perform two shows.</td></tr>
<tr><td>Selected Answer</td><td colspan="2">Andros Townsend's call-up</td><td>the band</td></tr>
<tr><td>Question Generation</td><td>BART-QA2D
What is Paul Merson not happy with to the England squad last week?</td><td>MixQG-large
What is Paul Merson not happy with?</td><td>BART-QA2D Question
Who will perform two shows?</td></tr>
<tr><td>QA Output</td><td>Townsend's call-up</td><td>he shouldn't have been in the squad</td><td>Unanswerable (Twisted Sister)</td></tr>
<tr><td>Answer Overlap</td><td>1.00</td><td>0.30</td><td>0.00 (0.80)</td></tr>
</table>

Table 4: Example source documents and summaries along with component outputs from the QA-based metric. *Left:* This example illustrates that the fluency of the QG model does not necessarily improve downstream factual consistency evaluation performance; the less fluent, more extractive BART-QA2D question is more-easily answerable by the QA model. Not shown in this table, the entailment-based SCConv metric incorrectly labels this entity-centric example, likely due the introduction of novel unigrams. *Right:* The QA model incorrectly labels this question as unanswerable, perhaps due to the generality of the question or due to noise in the input document. The QA output and score if forced to extract an answer are in parenthesis. SCConv correctly labels this highly extractive example.

the Answerability Penalty and using the answer overlap score for all questions, even those judged unanswerable by the QA model, also decreases performance. While the answer overlap metric should capture unanswerable questions for information not found in the input (extrinsic error), the answer from the answer selection component may appear in both the summary and source but in different contexts (intrinsic error). The QA model may return this answer and be scored as correct by the answer overlap component despite a factual inconsistency. This finding demonstrates the importance of determining question answerability, a point also emphasized in Deutsch et al. (2020) for QA-based metrics of relevance. Removing both of these components results in a drastic performance decrease.

5.2 Overall Results

We present the results on the test set of SummaC in Table 3. QAFactEval shows a substantial improvement over the previous state-of-the-art QA metric for factual consistency, QuestEval. Furthermore, it outperforms all other entailment-based metrics. QAFactEval-NLI shows slight im-

Model Type	Model Name	XSF	SummEval	FRANK-CNNDM	FRANK-XSum	QAGs-CNNDM	QAGs-XSum
Misc	BARTScore	0.25	0.37	0.58	0.15	**0.73**	0.17
	BLANC	0.03	0.20	0.33	0.07	0.33	0.02
Entailment	FactCC	0.04	0.37	0.38	0.06	0.40	0.30
	BertScore-FFCI	**0.45**	0.27	0.36	0.16	0.53	0.21
	DAE	0.02	0.45	0.50	0.22	0.63	-0.20
	ANLI	0.16	0.43	0.53	0.18	0.65	0.39
	MNLI	0.18	0.44	0.52	0.18	0.66	0.35
	DocNLI	0.01	0.41	0.12	0.26	0.16	-0.34
	SCZeroShot	0.06	0.50	0.55	0.27	0.57	**0.44**
QA	QuestEval	**0.45**	0.41	0.52	0.24	0.51	0.23
	QAFactEval	0.29	**0.61**	**0.66**	**0.32**	**0.68**	**0.44**
Learned	SCConv (synthetic)	0.12	0.50	0.59	**0.30**	0.03	0.06
	QAFactEval-NLI(synthetic)	0.19	**0.61**	**0.66**	0.25	0.65	**0.48**

Table 5: Instance-level Pearson correlation coefficients across factual consistency evaluation datasets. Metrics are divided into entailment-based, QA-based, and learned metrics that are fine-tuned on synthetic or supervised data. The two highest-correlated metrics for each dataset are shown in bold.

provements on the *synthetic* data. Notable improvements in this synthetic setting can be observed on the FactCC dataset, likely as the synthetic FactCC data the model is trained on was designed to mirror the errors captured in annotations. This performance boost on FactCC motivated our use of *supervised* data for fine-tuning our learned metric. Supervised fine-tuning on validation data helps in most cases and QAFactEval-NLI (supervised) improves on the overall benchmark by a statistically significant margin, using bootstrap resampling (Efron, 1982) with Bonferroni correction (Bonferroni, 1935) to obtain 99% confidence intervals (see Appendix for details). The performance drop on FactCC could be due to the proximity of the synthetic data to the labeled data and the data size difference. BertScore-FFCI performs best on XSF perhaps due to the closeness between BertScore's token-level metric and XSF's word-level annotations, and DocNLI's Polytope performance may also be from training data similarity.

We find that QAFactEval and SCConv do offer complementary signals that can be learned from supervised data. Individually fine-tuning the learned SCConv or a learned variation of QAFactEval on supervised data did not improve results over the non-supervised metrics; this result suggests the necessity of combining the two for further improvements. Training on the validation sets combined, rather than on each individual dataset separately, did not give an improvement, likely due to the learnable combination of NLI and QAFactEval being dataset dependent.

5.3 Correlation Analysis

We provide instance-level Pearson correlation between aggregated human judgments and metric scores for each model to compare to previous work in factual consistency that reports correlation analysis. Results are shown in Table 5. We split FRANK into CNN/DailyMail and XSum subsets for finer-grained analysis, as substantial differences have been noted in correlation performance across the two datasets (Durmus et al., 2020). We exclude Polytope, FactCC, and CGS here as prior work has only studied these datasets for binary classification.

We find that QAFactEval performs well across most datasets. As in the classification results, BertScore-FFCI's performs well on XSF, and we note that QuestEval's answerability classifier correlates more so with these fine-grained annotations than on other datasets. QAFactEval-NLI performs well on most datasets except XSF. Fine-tuning on FactCC synthetic data for binary classification likely does not capture the aggregated, word-level factuality scores of XSF. We leave a study of fine-tuning this model on supervised data with a regression loss for future work.

6 Conclusion

In this work, we demonstrated that QA-based metrics, when its components are properly optimized, outperform entailment-based metrics on a comprehensive factual consistency evaluation benchmark. We identify question generation and answerability detection as key components for improving QA-based metrics in future work. Furthermore, we show that entailment and QA-based metrics offer complementary signals through a combined metric that achieves state-of-the-art performance on this benchmark. We believe that our work lays the foundation for future work in QA-based metrics for factual consistency by offering a fairer comparison to other metrics across datasets and settings.

7 Ethical Considerations

Dataset Biases The underlying models of the metrics presented in this work are trained on documents in English and thus mainly represent the culture of the English-speaking populace. Political or gender biases may also exist in the datasets, and models, and subsequently the metrics, trained on these datasets may propagate these biases. We did not stress test these metrics for such biases and request that the users of these metrics be aware of these potential issues in applying them.

Misuse Potential and Failure Mode When properly used, the metrics described in this paper can be a useful tool for detecting summarization model errors. However, the current metrics fail to detect all factual inconsistencies, which must be remembered when applying these metrics as a filter for downstream applications. Factual inconsistencies in summaries could contribute to misinformation on the internet.

Environmental Cost The experiments described in the paper primarily make use of A100 GPUs. Most of the metrics have already been trained, in which case we simply ran inference using the existing models. We typically used a single GPU per experiment. Training learned answer overlap components can take a couple of hours, while experiments for learned metrics on SummaC take less than 10 minutes. These are the base models used in these experiments, with the number of parameters, in millions, in parentheses: BERT-base (110), BART-large (400), Electra-base (110), Electra-large (335), RoBERTa-large (355), T5-base (220), T5-large (770). Future work may analyze the effect of using distilled backbone models on factual consistency evaluation.

References

Mario Barrantes, Benedikt Herudek, and Richard Wang. 2020. Adversarial nli for factual correctness in text summarisation models. *ArXiv preprint*, abs/2005.11739.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Carlo E Bonferroni. 1935. Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni*, pages 13–60.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *ArXiv preprint*, abs/1809.02922.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2020. Towards question-answering as an automatic metric for evaluating the content quality of a summary.

Daniel Deutsch and Dan Roth. 2020. SacreROUGE: An open-source library for using and developing summarization evaluation metrics. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 120–125, Online. Association for Computational Linguistics.

Daniel Deutsch and Dan Roth. 2021. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Bradley Efron. 1982. *The Jackknife, the bootstrap and other resampling plans*. CBMS-NSF regional conference series in applied mathematics. SIAM, Philadelphia, PA. Lectures given at Bowling Green State Univ., June 1980.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Dan Friedman, Ben Dodge, and Danqi Chen. 2021. Single-dataset experts for multi-dataset qa. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Robin Jia, Mike Lewis, and Luke Zettlemoyer. 2021. Question answering infused pre-training of general-purpose contextualized representations.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2021. Ffci: A framework for interpretable automatic evaluation of summarization.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020. FFCI: A framework for interpretable automatic evaluation of summarization. *ArXiv preprint*, abs/2011.13662.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and Soumith Chintala. 2020. Pytorch distributed: Experiences on accelerating data parallel training. *Proc. VLDB Endow.*, 13(12):3005–3018.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Klaus-Michael Lux, Maya Sappelli, and Martha Larson. 2020. Truth or error? towards systematic analysis of factual errors in abstractive summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 1–10, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Lidiya Murakhovs'ka, Chien-Sheng Wu, Tong Niu, Wenhao Liu, and Caiming Xiong. 2021. Mixqg: Neural question generation with mixed answer types.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Gallinari Patrick, Lamprier Sylvain, Piwowarski Benjamin, Staiano Jacopo, and Wang Alex. 2021. Questeval: Summarization asks for fact-based evaluation. *ArXiv preprint*, abs/2103.12693.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020b. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv preprint*, abs/1910.03771.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Additional Data and Model Details

In this section, we provide details regarding statistical testing, benchmark statistics, and miscellaneous details regarding our QA-based experiments.

A.1 Statistical Testing

To determine whether the improvements on the SummaC benchmark are statistically significant, we perform significance tests using bootstrap resampling (Efron, 1982), following Laban et al. (2021). We compare our best model to the best-performing model from prior work on a given subset of the benchmark. We compare confidence intervals at significance levels of 0.05 and 0.01 and apply the Bonferroni correction (Bonferroni, 1935). Statistically significant differences at the 0.01 level exist between QAFactEval-NLI (supervised) and the best prior work on the FRANK subset and for the overall benchmark result. We do not see statistically significant differences on the other datasets in the benchmark. However, the statistically significant difference at the overall benchmark is notable; while other metrics may perform comparably or better on a given dataset, our metric demonstrates consistent good performance across datasets.

Dataset	# Valid	# Test	% Positive
CGS	1281	400	49.7
XSF	996	996	9.4
Polytope	634	634	87.2
FactCC	931	503	85.8
SummEval	850	850	90.6
FRANK	671	1575	33.2

Table 6: Statistics of the six datasets in the SummaC benchmark. We provide the number of validation and test set examples and the percentage of positive examples in the validation set.

A.2 Benchmark Statistics

For completeness, we provide additional statistics for the SummaC benchmark in Table 6. Due to the exclusion of Omission and Addition as factual consistency errors in the Polytope dataset, our dataset contains benchmark replication contains many more positive examples for that dataset. For XSF, we restrict the dataset to those examples with labels for factual consistency with respect to the source, as opposed to more general factuality labels which take into account world knowledge, which results in fewer examples than the original SummaC benchmark. This is the same subset as was used in Koto et al. (2021).

Please see the following links for the licenses of the datasets and annotations: CGS[2], XSF[3], FactCC[4], SummEval[5]. We did not find licenses for the remaining datasets analyzed in our study. The intended uses of these licenses align with our use for research purposes.

A.3 Model Parameters

Ablation experiments started from a combination that provided good initial validation results and then swapped components. Running every combination of QA-based metric components is expensive. We experimented with running an ablation of the QA models with a 2nd-best performing answer selection component *ALL*. This reduced all scores compared to using the NP Chunks component. This

[2] https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2002

[3] https://github.com/google-research-datasets/xsum_hallucination_annotations#license

[4] https://github.com/salesforce/factCC/blob/master/LICENSE.txt

[5] https://github.com/Yale-LILY/SummEval/blob/master/LICENSE

experiment supports our setup of keeping the best component constant when running ablations in order to determine the highest-performing combination of components, rather than experimenting with every combination.

Inference for the MADE QA model is run using the average of the six MADE adapters' parameters.

For Question Filtering with the IsAnsweredSumm Filter, in addition to if the Electra-large QA model labels the question as unanswerable, if the *F1* overlap score between the selected answer and the QA model output is less than 0.60, we remove this question. This filter was added only to IsAnsweredSumm and not IsAnsweredInput as answering questions based on the summary, from which the question was generated, should be an easy task. We reached this threshold based on a qualitative analysis of model outputs, although this number could have also been further tuned on the validation set.

B Additional Correlation Results

We provide additional correlation coefficients as a point of reference for future work. Instance-level correlations calculate the correlation between all instances, while the summary-level correlation computes the correlation between scores for each summary of the same input and then averages over inputs. Summary-level correlations are excluded for QAGS as this dataset does not contain annotations for multiple models, which is necessary to compute this score.

Model Type	Model Name	XSF	SummEval	FRANK-CNNDM	FRANK-XSum	QAGs-CNNDM	QAGs-XSum
Misc	BARTScore	0.25	0.34	**0.54**	0.14	**0.68**	0.17
	BLANC	0.07	0.20	0.33	0.06	0.30	0.03
Entailment	FactCC	0.05	0.37	0.41	0.05	0.49	0.26
	BertScore-FFCI	**0.45**	0.26	0.34	0.15	0.50	0.20
	DAE	0.00	0.40	0.49	0.20	0.58	-0.14
	ANLI	0.18	0.35	0.46	0.08	0.60	0.36
	MNLI	0.16	0.39	0.49	0.11	0.61	0.35
	DocNLI	0.01	0.34	0.11	0.21	0.21	-0.38
	SCZeroShot	0.06	0.39	0.48	0.23	0.52	0.44
QA	QuestEval	**0.43**	0.33	0.47	**0.24**	0.45	0.24
	QAFACTEVAL	0.30	**0.43**	**0.54**	**0.26**	**0.64**	**0.44**
Learned	SCConv (synthetic)	0.19	0.41	**0.54**	0.22	0.04	0.04
	QAFACTEVAL-NLI(synthetic)	0.16	**0.47**	**0.60**	0.21	**0.64**	**0.47**

Table 7: Instance-level Spearman correlation coefficients across factual consistency evaluation datasets. Metrics are divided into entailment-based, QA-based, and learned metrics that are fine-tuned on synthetic or supervised data. The two highest-correlated metrics for each dataset are shown in bold.

Model Type	Model Name	XSF	SummEval	FRANK-CNNDM	FRANK-XSum	QAGs-CNNDM	QAGs-XSum
Misc	BARTScore	0.17	0.27	0.42	0.12	**0.55**	0.14
	BLANC	0.05	0.15	0.25	0.05	0.24	0.02
Entailment	FactCC	0.03	0.29	0.31	0.04	0.38	0.21
	BertScore-FFCI	**0.31**	0.20	0.25	0.12	0.39	0.16
	DAE	0.00	0.32	0.38	0.16	0.47	-0.11
	ANLI	0.12	0.28	0.36	0.07	0.48	0.30
	MNLI	0.11	0.31	0.38	0.09	0.49	0.28
	DocNLI	0.01	0.27	0.08	0.17	0.17	-0.31
	SCZeroShot	0.04	0.31	0.37	0.18	0.41	0.36
QA	QuestEval	**0.30**	0.26	0.36	**0.20**	0.35	0.20
	QAFACTEVAL	0.22	**0.34**	**0.43**	**0.23**	**0.51**	**0.36**
Learned	SCConv (synthetic)	0.13	0.33	0.42	0.18	0.03	0.03
	QAFACTEVAL-NLI(synthetic)	0.11	**0.37**	**0.47**	0.17	**0.51**	**0.38**

Table 8: Instance-level Kendall correlation coefficients across factual consistency evaluation datasets. Metrics are divided into entailment-based, QA-based, and learned metrics that are fine-tuned on synthetic or supervised data. The two highest-correlated metrics for each dataset are shown in bold.

Model Type	Model Name	XSF	SummEval	FRANK-CNNDM	FRANK-XSum
Misc	BARTScore	0.18	0.40	0.65	0.29
	BLANC	0.12	0.27	0.47	0.01
Entailment	FactCC	-0.02	0.39	0.40	-0.07
	BertScore-FFCI	0.21	0.37	0.44	0.19
	DAE	0.01	0.51	0.54	0.32
	ANLI	0.09	0.49	0.53	0.18
	MNLI	0.10	0.48	0.51	0.17
	DocNLI	0.00	0.52	0.21	0.47
	SCZeroShot	0.11	0.57	0.60	**0.52**
QA	QuestEval	**0.30**	0.45	0.54	0.44
	QAFACTEVAL	**0.24**	**0.64**	**0.68**	**0.53**
Learned	SCConv (synthetic)	0.17	0.54	0.60	0.46
	QAFACTEVAL-NLI(synthetic)	0.16	**0.64**	**0.70**	0.48

Table 9: Summary-level Pearson correlation coefficients across factual consistency evaluation datasets. Metrics are divided into entailment-based, QA-based, and learned metrics that are fine-tuned on synthetic or supervised data. The two highest-correlated metrics for each dataset are shown in bold.

Model Type	Model Name	XSF	SummEval	FRANK-CNNDM	FRANK-XSum
Misc	BARTScore	0.18	0.38	**0.59**	0.28
	BLANC	0.12	0.25	0.43	0.06
Entailment	FactCC	0.00	0.37	0.42	-0.01
	BertScore-FFCI	0.21	0.34	0.40	0.20
	DAE	0.00	0.40	0.47	0.30
	ANLI	0.10	0.39	0.47	0.17
	MNLI	0.08	0.38	0.48	0.15
	DocNLI	-0.02	0.39	0.19	0.41
	SCZeroShot	0.11	0.41	0.51	**0.50**
QA	QuestEval	**0.27**	0.35	0.47	0.45
	QAFACTEVAL	**0.22**	**0.45**	**0.59**	0.47
Learned	SCConv (synthetic)	0.16	0.43	0.55	0.44
	QAFACTEVAL-NLI(synthetic)	0.17	**0.47**	**0.63**	**0.49**

Table 10: Summary-level Spearman correlation coefficients across factual consistency evaluation datasets. Metrics are divided into entailment-based, QA-based, and learned metrics that are fine-tuned on synthetic or supervised data. The two highest-correlated metrics for each dataset are shown in bold.

Model Type	Model Name	XSF	SummEval	FRANK-CNNDM	FRANK-XSum
Misc	BARTScore	0.15	0.32	**0.51**	0.25
	BLANC	0.11	0.21	0.38	0.05
Entailment	FactCC	0.00	0.30	0.35	-0.01
	BertScore-FFCI	0.17	0.28	0.34	0.18
	DAE	0.00	0.33	0.41	0.27
	ANLI	0.08	0.32	0.41	0.16
	MNLI	0.07	0.31	0.41	0.14
	DocNLI	-0.01	0.32	0.17	0.37
	SCZeroShot	0.10	0.34	0.44	**0.45**
QA	QuestEval	**0.23**	0.29	0.41	0.41
	QAFACTEVAL	**0.19**	**0.37**	0.51	**0.45**
Learned	SCConv (synthetic)	0.14	0.36	0.49	0.41
	QAFACTEVAL-NLI(synthetic)	0.14	**0.39**	**0.55**	0.44

Table 11: Summary-level Kendall correlation coefficients across factual consistency evaluation datasets. Metrics are divided into entailment-based, QA-based, and learned metrics that are fine-tuned on synthetic or supervised data. The two highest-correlated metrics for each dataset are shown in bold.