TRUE: Re-evaluating Factual Consistency Evaluation

Anonymous ACL submission

Abstract

Grounded text generation systems often generate text that contains factual inconsistencies, hindering their real-world applicability. Automatically evaluating such inconsistencies may help to alleviate this limitation by accelerating evaluation cycles, filtering inconsistent outputs and annotating large-scale training data. While attracting increasing attention, such evaluation metrics are usually developed and evaluated in silo for a single task or dataset. Moreover, previous meta-evaluation protocols focused on system-level correlations with human annotations, which leave the example-level accuracy of such metrics unclear. In this work, we introduce TRUE: a comprehensive study of factual consistency metrics on a standardized collection of existing texts from diverse tasks, manually annotated for factual consistency. Our standardization enables an example-level meta-evaluation protocol that is more actionable and interpretable than previously reported correlations, yielding clearer quality measures. Across diverse state-of-the-art metrics and 11 datasets we find that large-scale NLI and question generation-and-answering-based approaches achieve strong and complementary results, and recommend them as a starting point for future evaluations.¹

1 Introduction

A core issue in deploying text generation models for real-world applications is that they often generate factually inconsistent text with respect to the input they are conditioned on, or even completely “hallucinate” (Lee et al., 2018; Rohrbach et al., 2018; Maynez et al., 2020; Zhao et al., 2020) as exemplified in Table 1.

| Table 1: Factual inconsistencies (in red) from various tasks which are part of the TRUE study. The corresponding parts in the input/grounding are in blue. |
| --- |
| **Summarization (Wang et al., 2020)** |
| **Input** | Phyllis schlafly, a leading figure in the us conservative movement, has died at her home in missouri, aged 92... |
| **Summary** | Us conservative activist phyllis schlafly has died at the age of 87. |
| **Fact Verification (Thorne et al., 2018)** |
| **Evidence** | Ronald Bilius ”Ron” Weasley is a character in J. K. Rowling’s Harry Potter fictional series. |
| **Claim** | Ron Weasley is a President. |
| **Paraphrasing (Zhang et al., 2019)** |
| **Input** | The tracks were produced by Tommy Lee, and feature Michael Beinhorn on drums. |
| **Paraphrase** | The tracks were produced by Michael Beinhorn and have Tommy Lee on drums. |
| **Knowledge-Grounded Dialogue (Honovich et al., 2021)** |
| **Knowledge** | The first flip trick called a kickflip, originally called a ”magic flip,” was invented by professional skateboarder Rodney Mullen. |
| **Response** | I remember the first one was called magic flip. It was called a magic flip and was invented in the 60’s. |

¹Our code will be made publicly available.

increasing attention (Zhou et al., 2021) as they enable both better evaluation and better generation models via filtering training data (Gehrmann et al., 2021) or annotation of training data for controlled generation (Rashkin et al., 2021b).

While automatically evaluating factual consistency is an active line of work, there is no single agreed-upon meta-evaluation protocol for measuring the quality of such methods, and labeling schemes vary in their granularity. Works are usually done in silo, introducing new datasets and methods that target a specific task or domain, such as summarization (Falke et al., 2019; Kryscinski et al., 2020; Wang et al., 2020; Scialom et al., 2021; Deutsch et al., 2021; Xie et al., 2021) or dialogue
We show that Natural Language Inference (NLI) and Question Answering (QG-QA) approaches achieve significant improvement. Finally, we perform both quantitative and qualitative analysis of our results, finding that all approaches struggle with long inputs, pointing at challenges like long inputs and personal statements to be addressed in future work.

2 Standardizing Factual Consistency

In this section we elaborate on our re-evaluation setup. We first formally define what factual consistency refers to in this work. We then detail the datasets we consider and how we standardize them. Finally, we discuss the meta-evaluation protocol we propose for measuring the performance of evaluation methods on the standardized datasets.

2.1 Definitions and Terminology

We define a text to be factually consistent w.r.t its grounding text if all the factual information it conveys is consistent with the factual information conveyed by the grounding text. While some previous works distinguished between inconsistent erroneous text to inconsistent correct text (Maynez et al., 2020), we take a strict approach, requiring the text to be faithful to its grounding text, regardless of the “correctness” w.r.t the “real world”. In other words, we consider only the information present in the input text, not external knowledge, to assess faithfulness. This enables a more well-defined task, since determining the truthfulness of a fact w.r.t a general “real world” is subjective and depends on the knowledge, values and beliefs of the subject (Heidegger, 2001). This definition follows similar strictness in Textual Entailment, Question Answering, Summarization and other tasks where comprehension is based on a given grounding text, irrespective of contradiction with other world knowledge. This is also in line with recent work on evaluating attribution in text generation (Rashkin et al., 2021a), where humans are required to judge whether a generated text is true according to a grounding text. We use the terms consistent,
grounded, faithful and factual interchangeably.

### 2.2 Standardization Process

We include 11 datasets that contain human annotations w.r.t factual consistency in diverse tasks (Table 2). Other than the importance of covering a wide variety of error types, this also alleviates issues of rating quality which may vary across datasets (Denton et al., 2021).

To allow a unified evaluation framework we convert all annotations to binary labels that correspond to whether the entire target text is factual w.r.t the given grounding text or not. We note that a fine-grained annotation scheme, i.e., a typology of errors, was proposed for factual consistency (Pagnoni et al., 2021). While useful, most existing datasets do not include such labels. Moreover, while Machine Translation (MT) evaluation also showed value in fine-grained annotations (Freitag et al., 2021), it was proposed after years of improving MT to the level where coarse-grained annotation is insufficient. We argue that current grounded generation models are still at early stages w.r.t factual consistency, and that binary labeling is more beneficial now as it enables easier standardization across tasks and domains, with the goal of bringing researchers to collaborate on a shared methodology.

Binary annotation also corresponds to practical applications where filtering out unfaithful predictions is desired, and is in-line with the recommendations for human evaluation of attribution in text generation by Rashkin et al. (2021a).

We next detail the 11 datasets included in TRUE.

#### 2.2.1 Abstractive Summarization

**FRANK**  
Pagnoni et al. (2021) proposed a typology of factual errors, grounded in frame semantics (Fillmore, 1976; Palmer et al., 2005) and linguistic discourse theory (Brown and Yule, 1983). Based on this typology, they collected annotations for model-generated summaries on the CNN/DailyMail (CNN/DM; Hermann et al., 2015) and XSum (Narayan et al., 2018) datasets, resulting in 2250 annotated system outputs. Each summary sentence was annotated by three annotators. We take the majority vote for each sentence to get a sentence-level label and consider a summary as consistent if all sentences are consistent.

**SummEval**  
SummEval (Fabbri et al., 2020) is a comprehensive study of evaluation metrics for text summarization. The authors collected human judgments for 16 model outputs on 100 articles taken from the CNN/DM dataset, using both extractive and abstractive models. Annotators were asked to rate summaries on a Likert scale from 1 to 5, over 4 dimensions: consistency, coherence, fluency and relevance. Each summary was scored by 5 crowd-workers and 3 expert annotators. We label summaries as consistent only if all the expert annotators gave a consistency score of 5.

**MNBM**  
Maynez et al. (2020) annotated system outputs for the XSum dataset (Narayan et al., 2018). They sampled 500 articles and annotated summaries generated by four different systems, as well as the gold summaries. Annotators were asked to assess whether the summary includes hallucinations. Judgments from three different annotators were collected for each document-summary pair. To convert to a binary-label format, we use the binary consistency decision of whether a summary contains no hallucinations, and assign a label by taking the majority vote of the three annotators.

**QAGS**  
Wang et al. (2020) collected judgments of factual consistency on generated summaries for CNN/DM and XSum. Annotators were presented with the summaries one sentence at a time, along with the article, and determined whether each sentence is factually consistent w.r.t the article. Each sentence was annotated by 3 annotators, using the majority vote as the final score. To convert to binary-label format, we consider a summary consistent only if all its sentences are consistent.

#### 2.2.2 Dialogue Generation

**BEGIN** (Dziri et al., 2021) is a dataset for evaluating groundedness in knowledge-grounded dialogue systems, in which system outputs should be consistent with a grounding knowledge provided to the dialogue agent. BEGIN frames the task as NLI (Bowman et al., 2015), adopting the *entailment* and *contradiction* labels, and splitting the neutral

| Task          | # Examples | Open Test | Cons. |
|---------------|------------|-----------|-------|
| Summarization |            |           |       |
| - FRANK       |            |           |       |
| - SummEval    |            |           |       |
| - MNBM         |            |           |       |
| - QAGS        |            |           |       |
| - QAGS-DM     |            |           |       |
| Dialogue      |            |           |       |
| - BEGIN       |            |           |       |
| - Q3          |            |           |       |
| - DialFact    |            |           |       |
| Fact Verification |        |           |       |
| - FEVER       |            |           |       |
| - VitaminC    |            |           |       |
| Paraphrasing  |            |           |       |
| - PAWS        |            |           |       |

Table 2: Statistics for the datasets incorporated in TRUE. Cons. is the ratio of consistent examples.
label into three sub-categories: hallucination, off-topic responses and generic responses. Dialogue responses were generated by fine-tuning two systems on the Wizard of Wikipedia (WoW) dataset (Dinan et al., 2019), in which responses should be grounded in a span of text from Wikipedia. The generated responses were split into sentences, and each sentence was annotated separately. To convert to a binary-label format, we treat entailed sentences as consistent and all others as inconsistent.

$Q^2$ Honovich et al. (2021) annotated 1,088 generated dialogue responses for binary factual consistency w.r.t the knowledge paragraph provided to the dialogue model, for two dialogue models trained on WoW. Responses were annotated using binary labels by 3 of the paper authors, one annotator per response. We use $Q^2$’s labels without changes.

DialFact Gupta et al. (2021) introduced the task of fact-verification in dialogue, and constructed a dataset of conversational claims paired with pieces of evidence from Wikipedia. They define three tasks: (1) detecting whether a response contains verifiable content (2) retrieving relevant evidence and (3) predicting whether a response is supported by the evidence, refuted by the evidence or if there is not enough information to determine. We use the verifiable (i.e., factual, rather than personal) responses annotated for the third task, treating supported annotations as consistent and the rest as inconsistent. In cases where several evidence were marked as required for verification, we concatenate all evidence sentences to be the grounding text.

2.2.3 Fact Verification

FEVER Thorne et al. (2018) introduced FEVER (Fact Extraction and VERification), a dataset for fact verification against textual sources. FEVER was constructed by extracting information from Wikipedia, generating claims from it using annotators, then classifying whether each claim is supported or refuted by Wikipedia. Claims can also be labeled with NotEnoughInfo, meaning that there is not enough information in Wikipedia to either verify or refute the claim. Given a claim, the task defined by FEVER is to first extract evidence, then to determine whether it supports or refutes the claim. In a slightly different framing, the latter stage in FEVER is to determine whether the claim is factually consistent or not w.r.t the evidence, which is aligned with what we measure in TRUE. We use the development set of the NLI version of FEVER (Nie et al., 2019, 2020), treating supported claims as consistent and the rest as inconsistent.

VitaminC Schuster et al. (2021) derived a large-scale fact verification dataset from factual revisions to Wikipedia pages. Each example includes an evidence text from Wikipedia and a fact, with an annotation of whether the fact is supported, refuted or neutral w.r.t the evidence. The authors collected factual revisions to Wikipedia articles (pairs of “before” and “after” sentences), and asked annotators to write two facts for each pair: one that is supported by the first sentence and refuted by the second, and vice versa. When no explicit contradiction was present, the annotators wrote facts that are neutral w.r.t the evidence. Additional examples were created by revising examples from FEVER. We treat examples that include supported facts as consistent, and refuted or neutral facts as inconsistent.

2.2.4 Paraphrase Detection

PAWS Zhang et al. (2019) constructed a dataset for paraphrase identification with 108,463 paraphrase and non-paraphrase pairs with high lexical overlap, generated by controlled word swapping and back-translation, followed by judgments from human raters. Source sentences were drawn from Wikipedia and the Quora Question Pairs (QQP) corpus. We only use the examples with Wikipedia source sentences and view the binary paraphrase labels as consistency labels. We note that the definition of paraphrase is not equivalent to the definition of factual consistency, as a subset of a source text is not a paraphrase but may still be factually consistent with the source. However, PAWS was constructed such that non-paraphrases usually have contradicting meanings and is therefore relevant.

2.3 Meta-Evaluation

Previous work on evaluating factuality focused on measuring correlation with human judgments (Pagnoni et al., 2021). However, such numbers are not very informative when one is interested in evaluating the absolute performance of inconsistency detection methods that perform a binary decision w.r.t each input.

To conduct a more fine-grained evaluation at the single example level, we report the Receiver Operating Characteristic Area Under the Curve (ROC AUC) w.r.t binary detection of inconsistent examples. This is equivalent to AUC w.r.t consistency detection.
true positive rate (TPR, a.k.a. the recall) against
the false positive rate (FPR, a.k.a. the fallout) at
different possible thresholds for each tested metric.
Measuring ROC AUC evaluates the different met-
rics without setting a specific decision threshold.

For datasets with existing development/test split,
we also tune a threshold for the binary con-
sistency/inconsistency decision on the develop-
ment set and report the test set accuracy using
this threshold. We tune the thresholds by opti-
mizing the geometric mean of TPR and 1-FPR:
\[ \sqrt{\text{TPR}} \times (1 - \text{FPR}) \]

3 Evaluation Metrics

We compare various standard as well as state-of-
the-art approaches that measure factual consistency.
This comparison should draw a clear picture of cur-
rent research on this subject and directions for fu-
ture work. For example, we expect that robust met-
rics should perform well across tasks and datasets.
We next describe the different metrics tested as part
of this study. We note that for all reference-based
metrics, the grounding text serves as the reference.
For metrics where the scores are not in the [0,1]
range, we normalize scores to be in that range.

3.1 N-Gram Based Metrics

Standard N-Gram matching metrics such as
BLEU (Papineni et al., 2002) ROUGE (Lin, 2004)
and token-level F1 were shown to have weak cor-
relation with factual consistency (Maynez et al.,
2020; Honovich et al., 2021), with no exception on
TRUE. For completeness, we report their perfor-
ance in Table 9 in the appendix.

3.2 Model-Based Metrics

BERTScore (Zhang et al., 2020) aggregates simi-
larity scores between the BERT contextual embed-
ding of tokens in candidate and reference sentences.
We report results for the BERTscore-precision vari-
ant as it showed better results in preliminary exper-
iments. We use BERTScore version 0.3.11. with the
DeBERTa-xl-MNLI model (He et al., 2021; Nangia
et al., 2017), which is the recommended model as
of the time of writing this paper.6

BLEURT (Sellam et al., 2020a,b) is a learned met-
ric based on BERT (Devlin et al., 2019) for
evaluating text generation. BLEURT includes ad-
ditional pretraining on synthetic data followed by
fine-tuning on human judgements to train a model
that scores system outputs. We use the recom-
manded BLEURT-20 checkpoint (Pu et al., 2021).7

FactCC (Kryscinski et al., 2020) is a BERT-
based metric trained to verify factual consistency
of summaries. Training data was synthetically gen-
erated by applying rule-based transformations to
generate consistent and inconsistent summaries.

BARTScore (Yuan et al., 2021) evaluates text us-
ning probabilities from force-decoding with a BART
model (Lewis et al., 2020). We use the version fine-
tuned on the ParaBank2 dataset (Hu et al., 2019).

3.3 Natural Language Inference Metrics

ANLI The task of Textual Entailment (Dagan
et al., 2006) or Natural Language Inference (NLI;
Bowman et al., 2015) is to determine, given two
sentences, a hypothesis and a premise, whether the
hypothesis in entailed by the premise, contradicts it,
or is neutral w.r.t it. The resemblance of NLI to fac-
tual consistency evaluation has led to utilizing NLI
models for measuring factual consistency (Thorne
et al., 2018; Maynez et al., 2020; Dziri et al., 2021).
We trained an NLI model by fine-tuning T5-11B
(Raffel et al., 2020) on the Adversarial NLI (ANLI;
Nie et al., 2020) dataset. As suggested by Maynez
et al. (2020), we compute the entailment probabil-
ity with the grounding text as the premise and the
generated text as the hypothesis and use it as the
example-level factual consistency score.8

SUMMAC (Summary Consistency; Laban et al.,
2021) is focused on evaluating factual consistency
in summarization. They use NLI for detecting in-
consistencies by splitting the document and sum-
mary into sentences and performing NLI on all doc-
ument/summary sentence pairs, where the premise
is a document sentence and the hypothesis is a
summary sentence. They aggregate the NLI scores
for all pairs by either taking the maximum score
per summary sentence and averaging (SCZS) or
by training a convolutional neural network to ag-
gregate the scores (SCConv). We use the publicly
available implementation9 and report results for
SCZS as it performed better in our experiments.

---

6https://github.com/Tiiiger/bert_score
7https://github.com/google-research/bleurt/blob/master/checkpoints.md
8More implementation details on the NLI model are avail-
able in Section B in the appendix.
9https://github.com/tingofurro/summac
Table 3: ROC AUC results for the different metrics on the TRUE development set. We exclude VitaminC and FEVER from the average calculation as SC$_{ZS}$ was trained on VitaminC that includes examples from FEVER. The highest score in each row (excluding the Ensemble) is in bold and the aforementioned SC results are in strikethrough. Statistically significant results are indicated using * and ** for $p < 0.05$ and $p < 0.01$ respectively.

| Ensemble | Q$^2$ | ANLI | SC$_{ZS}$ | F1 | BLEURT | QuestEval | FactCC | BARTscore | BERTscore |
|----------|-------|------|-----------|----|--------|-----------|-------|-----------|-----------|
| FRANK   | 91.2  | 87.8 | 89.4      | 89.1| 76.4   | 82.8      | 84.0  | 76.4      | 86.1      | 84.3      |
| SummEval| 82.9  | 78.8 | 80.5      | 81.7| 61.4   | 68.7      | 70.1  | 75.9      | 75.5      | 71.2      |
| MNRM    | 76.6  | 68.7 | 77.3**    | 71.3| 46.2   | 64.5      | 65.3  | 59.4      | 60.9      | 62.1      |
| QAGS-C  | 87.7  | 80.5 | 83.8      | 78.1| 51.9   | 69.2      | 64.5  | 67.4      | 80.9      | 69.1      |
| QAGS-A  | 84.8  | 70.9 | 80.5      | 75.7| 57.2   | 63.6      | 64.9  | 53.8      | 49.3      |
| BEGIN   | 86.2  | 79.7 | 82.6      | 82.0| 86.4   | 84.1      | 64.4  | 86.3      | 87.9      |
| Q$^2$   | 82.8  | 80.9 | 72.7      | 77.4| 65.9   | 72.4      | 67.9  | 64.9      | 70.0      |
| DialFact| 96.5  | 86.7 | 84.1      | 72.3| 73.1   | 77.3      | 55.3  | 65.6      | 64.2      |
| PAWS    | 91.2  | 89.7**| 86.4  | 88.2| 51.4   | 68.3      | 69.2  | 64.0      | 71.5      | 71.5      |
| FEVER   | 94.7  | 88.4 | 93.2**    | 92.2| 51.8   | 59.5      | 72.6  | 61.9      | 64.1      | 63.3      |
| VitaminC| 96.1  | 81.4 | 88.3**    | 92.2| 61.4   | 61.8      | 66.5  | 56.3      | 63.2      | 62.5      |
| Avg. w/o V.C. FEVER | 88.0 | 80.7 | 81.5 | 81.4 | 63.8 | 71.4 | 71.4 |

3.4 QG-QA Based Metrics

Durmus et al. (2020) and Wang et al. (2020) proposed to use Question Generation (QG) and Question Answering (QA) models to automatically evaluate factual consistency in abstractive summarization, showing promising results. Honovich et al. (2021) employed a similar approach for evaluating knowledge-grounded dialogue generation.

The steps of the QG-QA approach are as follows: (1) Questions are automatically generated for spans in the generated text, such that the answer to a question is its respective input span. (2) The generated questions are answered using a QA model on the grounding text, resulting in an answer span or a “no-answer” output. (3) For each question, the two answer spans from the grounding and the generated text are compared to get a score. (4) The scores for all questions are aggregated into a final score.

**Q$^2$** (Honovich et al., 2021) is a QG-QA method that employs an NLI model to compare the two answers for each question, where the grounding text answer is the premise and the generated text answer is the hypothesis. We report results for a re-implementation of $Q^2$ using T5-11B as the backbone for the QA, QG and NLI models. While Honovich et al. (2021) validate each generated question by answering it using a QA model and comparing to the original extracted answer candidate using exact match, we relax this and instead use F1 token-overlap with a predefined threshold.\footnote{More implementation details are available in Section B in the appendix.}

**QuestEval** (Scialom et al., 2021) is a QG-QA method that measures both factual consistency and relevance (by reversing the roles of the generated and grounding texts). The authors trained a model that weighs each generated question according to the relevance of its answer to appear in the generated text. Their results showed high correlation with human judgments in comparison to prior work on the SummEval benchmark (Fabbri et al., 2021). We use the publicly available version.\footnote{https://github.com/ThomasScialom/QuestEval}

4 Results

We report the ROC AUC\footnote{Multiplied by 100 for better readability.} of various metrics on the standardized datasets in Table 3. The ROC curves can be found in Figure 2 in the appendix. As all metrics operate in a “zero-shot” manner on all datasets (except for SUMMAC on VitaminC and FEVER) and no threshold tuning is required, we report results on the development sets.\footnote{AUC and accuracy for the test sets are provided in Tables 10 and 11 in the appendix.} SC$_{ZS}$ was trained on VitaminC which includes examples from FEVER, so we exclude those datasets from the average AUC calculation for a more fair comparison.

The results show that the NLI-based models (ANLI, SC$_{ZS}$) outperformed the other approaches on 6 datasets, with average AUC of 81.5 and 81.4 for ANLI and SC$_{ZS}$, respectively. $Q^2$ outperform the other approaches on 4 datasets, with an average AUC of 80.7. The next best method, BARTScore, had lower average AUC of 72.2. All other approaches scored 72 or lower on average across all datasets (excluding FEVER and VitaminC).

One outlier is BEGIN, which is the only dataset where simple metrics like F1 token overlap achieved scores higher than 80. We measured the average overlap between the grounding and target texts per dataset, and found that BEGIN exhibits a high difference between grounded and ungrounded texts in comparison to other datasets (Table 8 in appendix A), which explains this.
We follow Laban et al. (2021) and perform significance testing through bootstrap resampling (Efron, 1982), comparing the best method to the second-best method on each dataset. We perform interval comparison at $p = 0.05$ and $p = 0.01$ and find significantly best results on 6 datasets, 3 from $Q^2$ and 3 from ANLI.

Given that no single method outperformed the rest on all datasets, we hypothesize that the NLI and QG-QA based metrics are complementary. We test this by averaging the $Q^2$, ANLI and SC$_{ZS}$ scores per example (Ensemble in Table 3). Indeed, averaging the three methods yields better results on most datasets and on average, with an increase of 4.5 in ROC AUC from the best single-metric result.

Our results show that a single metric can do well across all tasks and datasets, with all 3 best metrics scoring higher than 70 on all 11 datasets. This corroborates our hypothesis that evaluating factual consistency can be unified, and we hope such unified perspective will be adopted in future work to accelerate progress on the subject.

5 Analysis

Input Length. As QA and NLI models may struggle with long inputs (Kočiský et al., 2018; Pang et al., 2021; Yin et al., 2021; Shaham et al., 2022), metrics based on them may fail when handling long text. To study the effect of input length on the metrics performance, we unify all datasets and split examples into 6 bins according to the grounding length. We focus on the grounding as the target texts are usually short (see Table 6 in Appendix A). We measure AUC of the best 3 metrics according to their overall score for each length bin, sampling 1,000 examples per bin.

The results are shown in Figure 1. We find that there is a consistent degradation for texts longer than 200 tokens for all metrics, including SC$_{ZS}$ which is designed to better handle long text. We find it surprising that the ANLI-based model and $Q^2$ still do relatively well on the longest bin as they are required to perform end-to-end QA and NLI on texts with more than 500 tokens.

ModelSize. Model-based metrics are expected to benefit from increasing model size. To quantify this we study the effect of using smaller models for the ANLI, BLEURT and BERTScore metrics. We compare the average ROC AUC of larger and smaller model variants for each metric. We find an advantage of 4.7, 3.7 and 1.3 average ROC AUC for the larger ANLI, BLEURT and BERTScore variants respectively, showing that larger models are important for evaluating factuality. The full results are in Table 7 in the appendix.

Qualitative Analysis. We conduct manual error analysis to point at weaknesses of the different metrics and present challenges posed by the task. We analyze 80 examples that were misclassified by all three best metrics, as well as 100 examples that were correctly classified by one or two of the three.

Out of the analyzed examples, many seem to have a wrong label. This is especially true for cases in which all best metrics failed, with annotation errors in 35/80 cases. For the cases where one or two metrics failed, we found annotation errors in 27/100 cases. To verify that the high annotation error rate is indeed a result of inspecting the “hardest” examples, we uniformly sample additional 100 examples, finding that only 10 had annotation errors. This is in line with the findings of Freitag et al. (2021), who showed that in some cases, metrics may be better than non-expert annotators.

Despite showing impressive results, the best-performing metrics fail to detect subtle inconsistencies, as presented in Table 4. This was the case for 21/180 analyzed examples. Metrics that aggregate scores across parts of a target text, such as $Q^2$ or SC$_{ZS}$, might assign a high score for texts in which all but a small part is consistent. End-to-end NLI should predict “contradiction” even when only a small part of the text contradicts the grounding, but it may fail to do so. Applying a strict approach...
| Grounding                                                                 | Generated Text                                                                 | Explanation                                                                 |
|--------------------------------------------------------------------------|------------------------------------------------------------------------------|----------------------------------------------------------------------------|
| The word “philately” is the English version of the French word “philatélie”, coined by Georges Herpin in 1864. | The word philately is actually a fench word coined by george herpin.          | The word philately is an English word based on a French word, but not French. All best metrics misclassified this. |
| French police have interviewed presidential candidate Francois Fillon and his wife Penelope over claims she was paid for fake work. They provided information that would help find the "truth", mr Fillon said... | French presidential candidate Francois Fillon has said he and his wife penelope have been questioned by police over claims she worked illegally. | Most details are correct and the hallucination is subtle. In the case of $Q^2$, most of the generated questions have the same answer based on the grounding and the generated text, therefore the overall score was high. |
| Stamp collecting is generally accepted as one of the areas that make up the wider subject of philately, which is the study of stamps. | I've never heard of stamps, but I do know that the word "philately" refers to the study of stamps. | The personal statement “I’ve never heard of stamps” is not factual and should not be evaluated. |
| Evidence suggests that cognitive behavioral therapy and a gradual increase in activity suited to individual capacity can be beneficial in some cases. | It has been suggested that cognitive behavioral therapy and gradual increase in exercise could help in some cases so I’m going to try that for now. | Similar to the previous examples - SummaC, and ANLI falsely marked the text as inconsistent, probably due to the personal statement. |

Table 4: Examples for the error analysis. The first two rows show cases of challenging inconsistencies, while the last two show dialogue responses containing non-factual personal statements.

---

6 Related Work

Adding to the related work mentioned throughout the paper, works on unified evaluation of text generation across tasks include GEM (Gehrmann et al., 2021), where the focus is on evaluating system outputs and not the factuality evaluation methods as in TRUE. BEAMetrics (Scialom and Hill, 2021) proposes meta-evaluation protocols across tasks, but does not focus on factuality. When discussing factuality (“correctness”) they measure correlations, which are not sufficient as mentioned in Section 2.3. Other works on meta-evaluation of factuality across datasets include GO-FIGURE (Gabriel et al., 2021) FRANK (Pagnoni et al., 2021) and SummaC (Laban et al., 2021), however they all focus solely on summarization. To the best of our knowledge, our work is the first to generalize the discussion on evaluating factuality across tasks and datasets outside of summarization, and the first to show that large-scale QG-QA and NLI are highly complementary – setting stronger baselines for future work than previously published.

7 Discussion and Future Work

We discuss the main takeaways of the TRUE study, pointing at actionable insights for future work.

---

8 Conclusions

We presented TRUE, a meta-evaluation study for factual consistency. We standardized various datasets from diverse tasks into a unified labeling scheme to perform a thorough analysis of automatic evaluation methods, showing that NLI and QG-QA based approaches perform well across multiple tasks and datasets. We further show these methods are highly complementary – hinting at additional headroom for improvement while pointing on current limitations. We hope our results and methodology will encourage a more unified perspective in future work to foster progress towards more factual NLP applications.
References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Gillian Brown and George Yule. 1983. Discourse Analysis. Cambridge University Press.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Emily Denton, Mark Diaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. arXiv preprint arXiv:2112.04534.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. Transactions of the Association for Computational Linguistics, 9:774–789.

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.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In Proceedings of the International Conference on Learning Representations (ICLR).

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.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating groundedness in dialogue systems: The begin benchmark.

Bradley Efron. 1982. The jackknife, the bootstrap and other resampling plans. SIAM.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. arXiv preprint arXiv:2007.12626.

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.

C. Fillmore. 1976. Frame semantics and the nature of language *. Annals of the New York Academy of Sciences, 280.

Markus Freitag, George Foster, David Grangier, Vireesh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. arXiv preprint arXiv:2104.14478.

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.

Sebastian Gehrmann, Tosin Adewumi, Karmany Aggarwal, Pawan Sasanaka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyati Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenyem Emezue, Varun Gangal, Cristina Garbaceae, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khya Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Oseri, Ankur Parikh, Laura Perez-Beltrachini, Niranjani Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobel, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukokula, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), pages 96–120. Online. Association for Computational Linguistics.
Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Dialfact: A benchmark for fact-checking in dialogue. *arXiv preprint arXiv:2110.08222.*

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations.*

Martin Heidegger. 2001. On the essence of truth. *The Nature of Truth: Classic and Contemporary Perspectives,* 1:295–316.

Karl Moritz Hermann, Tomas Kocisky, 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,* volume 28. Curran Associates, Inc.

Or Honovich, Leshem Choshen, Roeo Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q2*: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL),* pages 44–54, Hong Kong, China. Association for Computational Linguistics.

Tomáš Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics,* 6:317–328.

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. *arXiv preprint arXiv:2111.09525.*

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942.*

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. *NeurIPS 2018 Workshop on Interpretability and Robustness for Audio, Speech, and Language.*

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining 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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out,* pages 74–81, Barcelona, Spain. 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 arXiv:1907.11692.*

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.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The RepEval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP,* pages 1–10, Copenhagen, Denmark. 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, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI).*

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.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual...*
Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1699–1713. Online. Association for Computational Linguistics.

Artidor 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.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics, 31(1):71–106.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, et al. 2021. Quality: Question answering with long input texts, yes! arXiv preprint arXiv:2112.08608.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuvan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1173–1186. Online. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 751–762. Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. 2021. Don’t be contradictory with anything! CI-ToD: Towards benchmarking consistency for task-oriented dialogue system. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2357–2367. Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2021a. Measuring attribution in natural language generation models. arXiv preprint arXiv:2112.12870.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021b. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 704–718. Online. Association for Computational Linguistics.

Ehud Reiter and Craig Thomson. 2020. Shared task on evaluating accuracy. In Proceedings of the 13th International Conference on Natural Language Generation, pages 227–231, Dublin, Ireland. Association for Computational Linguistics.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4035–4045.

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, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6594–6604. Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom and Felix Hill. 2021. Beametrics: A benchmark for language generation evaluation evaluation. arXiv preprint arXiv:2110.09147.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020a. BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7881–7892. Online. Association for Computational Linguistics.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020b. Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In Proceedings of the Fifth Conference on Machine Translation, pages 921–927. Online. Association for Computational Linguistics.
Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. Scrolls: Standardized comparison over long language sequences.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. 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.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. In Advances in Neural Information Processing Systems.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Zheng Zhao, Shay B Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. arXiv preprint arXiv:2009.13312.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1393–1404, Online. Association for Computational Linguistics.
A Additional Data Statistics

Tables 5 and 6 presents statistics regarding the length of the grounding text and the generated text for TRUE’s datasets, respectively.

![Table 5: Grounding length statistics for TRUE.](image)

![Table 6: Generated text length statistics for TRUE.](image)

Table 7 presents the results of an ablation study testing the effect of model size for different model-based metrics.

![Table 7: Ablation study comparing the average ROC AUC results for models with different sizes. “BERTScore P” stands for BERTScore Precision.](image)

B Implementation Details

We train all models using the t5x library.\(^{17}\)

QG-QA For our reimplementation of \(Q^2\) (Honovich et al., 2021) we use T5-11B as the pretrained model for QG, QA and NLI, while Honovich et al. (2021) used T5-Base, ALBERT (Lan et al., 2019), and RoBERTa (Liu et al., 2019) for the QG, QA and NLI models, respectively. We use a maximum length of 2048 tokens for the input. We set the F1 token overlap threshold to 0.54 by tuning it on a held-out dataset. We use beam search with a beam size of 4 to generate multiple questions, and use the first question that passes the validation threshold.

D ROC Curves

Figure 2 presents the ROC curves for the different datasets studied in TRUE, using the best-performing metrics.

---

\(^{17}\)https://github.com/google-research/t5x
### Table 8: Average overlap between the generated text and the grounding, measured using ROUGE-L and simple F1 token-overlap, taking the grounding to be the reference text. The “Pos” columns contain the statistics for the grounded text, while the “Neg” columns contain the statistics for the ungrounded text.

| Dataset     | Pos ROUGE_L | Neg ROUGE_L | ROUGE_L diff | Pos F1 | Neg F1 | F1 diff |
|-------------|-------------|-------------|--------------|--------|--------|---------|
| FRANK       | 0.105       | 0.060       | 0.045        | 0.165  | 0.103  | 0.062   |
| SummEval    | 0.181       | 0.141       | 0.041        | 0.282  | 0.244  | 0.038   |
| MNBM        | 0.044       | 0.047       | 0.003        | 0.079  | 0.084  | 0.006   |
| QAGS-CNNMDM | 0.215       | 0.170       | 0.045        | 0.281  | 0.249  | 0.031   |
| QAGS-XSUM   | 0.051       | 0.050       | 0.002        | 0.082  | 0.080  | 0.002   |
| BEGIN       | 0.465       | 0.159       | 0.306        | 0.553  | 0.207  | 0.346   |
| Q$^2$       | 0.228       | 0.169       | 0.059        | 0.368  | 0.264  | 0.104   |
| DialFact    | 0.302       | 0.200       | 0.102        | 0.394  | 0.249  | 0.144   |
| PAWS        | 0.832       | 0.734       | 0.098        | 0.938  | 0.934  | 0.003   |
| FEVER       | 0.174       | 0.179       | 0.005        | 0.276  | 0.258  | 0.018   |
| VitaminC    | 0.314       | 0.270       | 0.044        | 0.362  | 0.290  | 0.072   |

Figure 2: ROC curves for the best performing methods.
| Ensemble | Q^2 | ANLI+SC | BLEURT | QuestEval | FactCC | BARTscore | BERTscore |
|----------|-----|---------|--------|-----------|--------|-----------|-----------|
| FRANK    | 83.0| 81.5    | 82.0   | 79.0      | 76.6   | 73.0      | 72.4      | 80.7      | 75.6     |
| BEGIN    | 76.8| 74.1    | 76.8   | 78.9      | 74.3   | 73.4      | 62.09     | 74.8      | 78.1     |
| DialFact | 80.9| 78.1    | 68.4   | 74.2      | 67.7   | 69.0      | 72.5      | 78.0      | 60.2     |
| PAWS     | 84.8| 84.1    | 82.1   | 82.3      | 82.9   | 84.8      | 60.7      | 70.9      | 69.8     |
| VitaminC | 92.1| 77.5    | 83.9   | 59.0      | 63.3   | 55.5      | 59.8      | 58.0      |          |
| Avg.     | 81.4| 79.4    | 77.3   | 78.6      | 70.2   | 70.0      | 62.4      | 71.4      | 80.9     |

Table 9: ROC AUC results for metrics that were not reported in Table 3.

| Ensemble | Q^2 | ANLI | SC | BLEURT | QuestEval | FactCC | BARTscore | BERTscore |
|----------|-----|------|----|--------|-----------|--------|-----------|-----------|
| FRANK    | 89.6| 91.1 | 90.4| 88.9   | 80.1      | 76.0   | 80.1      | 76.0      |
| SummEval | 80.7| 83.0 | 82.0| 79.8   | 68.8      | 60.2   | 68.8      | 60.2      |
| MSMT8   | 75.6| 77.4 | 78.6| 87.2   | 47.5      | 69.3   | 47.5      | 69.3      |
| QA      | 86.0| 84.7 | 86.4| 79.6   | 67.1      | 63.9   | 67.1      | 63.9      |
| QA-M   | 81.8| 83.4 | 79.1| 76.1   | 52.9      | 48.6   | 52.9      | 48.6      |
| BEGIN   | 85.7| 82.4 | 85.7| 81.6   | 86.4      | 84.6   | 81.6      | 86.4      |
| Q^2     | 83.0| 76.9 | 83.9| 77.5   | 66.8      | 64.3   | 77.5      | 66.8      |
| DialFact| 89.4| 84.5 | 90.2| 81.2   | 71.2      | 72.5   | 81.2      | 71.2      |
| PAWS    | 90.5| 89.7 | 91.4| 88.2   | 82.2      | 77.3   | 82.2      | 77.3      |
| FEVER   | 90.3| 72.6 | 72.5| 20.1   | 39.9      | 71.1   | 39.9      | 71.1      |
| VitaminC| 79.3| 68.6 | 76.5| 29.8   |          | 59.6   | 29.8      |          |
| Avg.    | 84.7| 83.8 | 84.9| 80.0   | 69.2      | 66.5   | 80.0      | 69.2      |

Table 10: ROC AUC results for the different metrics on the TRUE test set. We exclude VitaminC from the average calculation as SC was trained on VitaminC. The highest score in each row (excluding the Ensemble) is in bold and the aforementioned SC results are in strikethrough.

| Ensemble | Q^2 | ANLI | SC | BLEURT | QuestEval | FactCC | BARTscore | BERTscore |
|----------|-----|------|----|--------|-----------|--------|-----------|-----------|
| FRANK    | 83.0| 81.5 | 82.0| 79.0   | 76.6      | 73.0   | 72.4      | 80.7      | 75.6     |
| BEGIN    | 76.8| 74.1 | 76.8| 78.9   | 74.3      | 73.4   | 62.09     | 74.8      | 78.1     |
| DialFact | 80.9| 78.1 | 68.4| 74.2   | 67.7      | 69.0   | 72.5      | 78.0      | 60.2     |
| PAWS     | 84.8| 84.1 | 82.1| 82.3   | 82.9      | 84.8   | 60.7      | 70.9      | 69.8     |
| VitaminC | 92.1| 77.5 | 83.9| 59.0   | 63.3      | 55.5   | 59.8      | 58.0      |          |
| Avg.     | 81.4| 79.4 | 77.3| 78.6   | 70.2      | 70.0   | 62.4      | 71.4      | 80.9     |

Table 11: Accuracy results for the different metrics on the TRUE test set. Thresholds were tuned on the corresponding development sets. We exclude VitaminC from the average calculation as SC was trained on VitaminC. The highest score in each row (excluding the Ensemble) is in bold and the aforementioned SC results are in strikethrough.