

TrueTeacher: Learning Factual Consistency Evaluation with Large Language Models

Zorik Gekhman^{T,G,*} Jonathan Herzig^G Roei Aharoni^G Chen Elkind^G Idan Szpektor^G

^TTechnion - Israel Institute of Technology ^GGoogle Research

zorik@campus.technion.ac.il

{zorik|jherzig|roeeaharoni|chenel|szpektor}@google.com

Abstract

Factual consistency evaluation is often conducted using Natural Language Inference (NLI) models, yet these models exhibit limited success in evaluating summaries. Previous work improved such models with synthetic training data. However, the data is typically based on perturbed *human-written* summaries, which often differ in their characteristics from real *model-generated* summaries and have limited coverage of possible factual errors. Alternatively, large language models (LLMs) have recently shown promising results in directly evaluating generative tasks, but are too computationally expensive for practical use. Motivated by these limitations, we introduce TrueTeacher, a method for generating synthetic data by annotating diverse *model-generated* summaries using a LLM. Unlike prior work, TrueTeacher does not rely on human-written summaries, and is multilingual by nature. Experiments on the TRUE benchmark show that a student model trained using our data, substantially outperforms both the state-of-the-art model with similar capacity, and the LLM teacher. In a systematic study, we compare TrueTeacher to existing synthetic data generation methods and demonstrate its superiority and robustness to domain-shift. Using the mFACE dataset, we also show that our method generalizes to multilingual scenarios. Finally, we release a large-scale synthetic dataset with 1.4M examples generated using TrueTeacher.¹

1 Introduction

Generative summarization models are prone to generate factually inconsistent summaries w.r.t. their input documents (Goodrich et al., 2019; Kryscinski et al., 2019). Since factual consistency evaluation could be cast as a Natural Language Inference

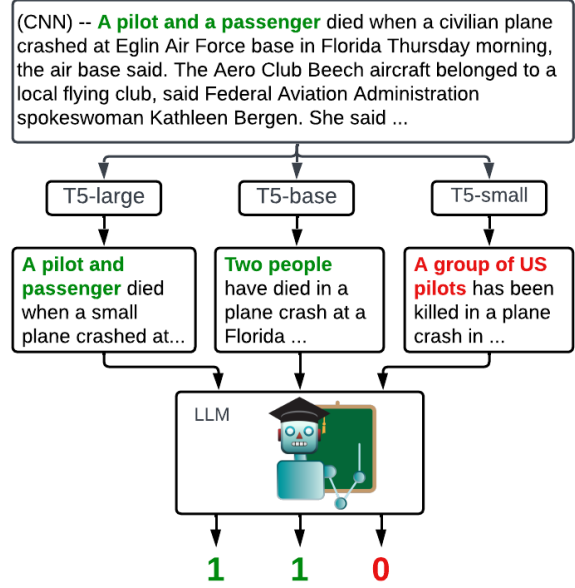


Figure 1: An example of our data generation process. We produce a diverse set of *model-generated* summaries of articles from CNN/DailyMail and label them for consistency using FLAN-PaLM 540B.

(NLI) task, NLI models are often used to detect such inconsistencies (Falke et al., 2019a; Maynez et al., 2020; Laban et al., 2022). However, NLI models exhibit limited success in evaluating factual consistency in *summarization* (Falke et al., 2019b; Kryscinski et al., 2020), since NLI datasets lack the entailment phenomena that naturally arise in abstractive summarization (Khot et al., 2018). For example, single-sentence premise-hypothesis pairs are shorter than document-summary pairs (Mishra et al., 2021; Schuster et al., 2022).

To adapt NLI models for evaluating document-summary pairs, previous work proposed to generate synthetic training data (Kryscinski et al., 2020; Yin et al., 2021; Utama et al., 2022; Balachandran et al., 2022). The data is typically generated by perturbing human-written summaries to introduce factual inconsistencies. While these perturbations are effective, they are limited to factual error categories that can be covered by the perturbation logic. Fur-

*Work done during an internship at Google Research.

¹https://github.com/google-research/google-research/tree/master/true_teacher

thermore, since the synthetic summaries are based on *human-written* summaries, they may differ in style from real *model-generated* summaries, which can further affect performance.

An alternative approach to augmenting NLI models with synthetic data, is to directly prompt large language models (LLMs) to evaluate factual consistency. Recently, there has been a growing evidence for the effectiveness of LLMs in evaluating generative tasks (Kocmi and Federmann, 2023; Wang et al., 2023; Liu et al., 2023), including factual consistency in summarization (Luo et al., 2023). However, LLMs are still too computationally expensive to be heavily used in practice.

To make the best of both worlds we propose TrueTeacher, a simple and effective synthetic data generation method that leverages *model-generated* summaries and the reasoning abilities of LLMs (Huang and Chang, 2022). In TrueTeacher (portrayed in Figure 1), we first train a diverse collection of summarization models with different capacities. Next, we use these models to summarize each document in a given corpus. The resulting document-summary pairs are then annotated by prompting a LLM to predict the corresponding factual consistency label.

We apply TrueTeacher using FLAN-PaLM 540B as the LLM (Chung et al., 2022) and generate a large-scale synthetic dataset that is used for training a student model. Experiments on the summarization subset of the TRUE benchmark (Honovich et al., 2022) show that adding the synthetic TrueTeacher data to existing NLI data improves the state-of-the-art single model’s ROC-AUC from 82.7 to 87.8. The resulting model even outperforms FLAN-PaLM, despite the latter being used as the teacher and having a $\times 50$ times larger capacity.

We also compare TrueTeacher to existing synthetic data generation methods. We notice that previous work vary in their generated data size, the corpus used for data synthesis, the model used for training, and the evaluation sets, which makes it difficult to directly compare between methods. Therefore, we design a systematic study to re-evaluate existing methods and discover that they generalize poorly when evaluated on documents from a different distribution than the one used to generate the synthetic data. Conversely, TrueTeacher successfully generalizes to documents from new domains, which demonstrates its robustness.

Finally, we apply TrueTeacher to generate syn-

thetic *multilingual* data. While synthetic data generation methods are often limited to English (Utama et al., 2022; Balachandran et al., 2022), TrueTeacher can easily utilize a multilingual LLM. Results on the mFACE dataset (Aharoni et al., 2022) for evaluating multilingual summaries using our method for multilingual data generation, show improvements on 35 out of 45 languages.

To summarize, this work includes the following contributions:

- We introduce TrueTeacher, a synthetic data generation approach based on annotating *model-generated* summaries with LLMs, and demonstrate its effectiveness and robustness.
- We evaluate FLAN-PaLM 540B on the task of factual consistency evaluation and show that its knowledge can be distilled into a significantly smaller model by utilizing our method.
- We conduct a systematic study, re-evaluating existing synthetic data generation methods for the task in an apples-to-apples comparison and identify their limitations.
- We perform the first experiment in generating multilingual synthetic data for factual consistency, and demonstrate its usefulness.
- We release a dataset with 1.4M TrueTeacher training examples, used in our experiments.

2 TrueTeacher

In this section we describe TrueTeacher, our approach for generating synthetic examples for the task of factual consistency evaluation in summarization. Our main motivation is to use factual inconsistencies from real *model-generated* summaries instead of perturbed gold-summaries. To this end, we generate a diverse set of summaries using generative summarization models of different capacities, and leverage a LLM to label them for consistency. Some of the output summaries are expected to contain consistency errors, and we hypothesize that a strong performing LLM can generalize to the task and label them with sufficient quality to be useful for training. The usage of model-generated summaries not only yields more realistic texts, but also allows to potentially include rare errors, which are harder to incorporate with perturbation logic.

Our data generation process is illustrated in Figure 2. To create *model-generated* summaries, we first train a variety of summarization models. We

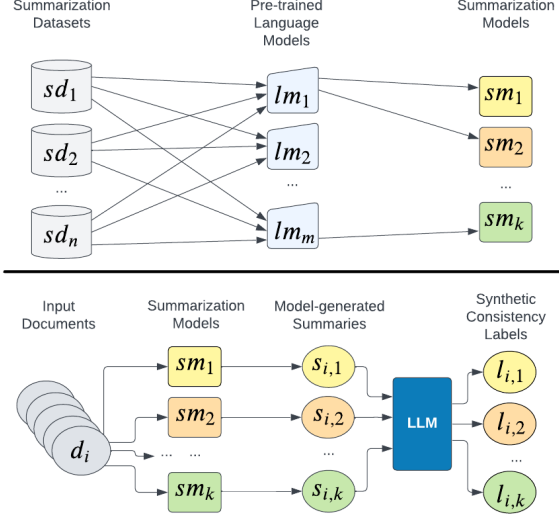


Figure 2: Our data generation process. We train a collection of generative summarization models, use them to summarize documents and label the resulting summaries for factual consistency using a LLM.

use a collection of one or more summarization training sets $T = \{sd_1, sd_2, \dots, sd_n\}$ and different pretrained $LMs = \{lm_1, lm_2, \dots, lm_m\}$ to fine-tune a collection of summarization models $SM = \{sm_1, sm_2, \dots, sm_k\}$, where $k = n \times m$.² Using different pretrained LMs allows to diversify the expected consistency errors, e.g., errors made by large or small models. The choice of summarization training sets allows to control for the nature of the resulting summaries, e.g., focusing on abstractive training sets to increase output abstractiveness.

Next, we choose a documents corpus $D = \{d_1, d_2, \dots, d_r\}$ and use all the summarization models in SM to summarize all the documents in D , resulting in a collection of model-generated output summaries $O = \{s_{1,1}, \dots, s_{r,k}\}$, where $s_{i,j}$ is the summary for document d_i generated by summarization model sm_j . TrueTeacher does not require gold summaries, which allows it to be used with any collection of documents D , and makes it more scalable than previous methods (Yin et al., 2021; Utama et al., 2022; Balachandran et al., 2022).

Finally, a LLM is prompted to label all the summaries in O for consistency w.r.t. their source documents, resulting with labels $\{l_{1,1}, \dots, l_{1,k}, \dots, l_{r,k}\}$.³ Bottom part of Figure 2 illustrates this process for a single document $d_i \in D$. Each document, summary, and label $(d_i, s_{i,j}, l_{i,j})$

²We note that the pretrained LMs here refer to the models that we are fine tuning for summarization, and they are different from the LLM that we use as the teacher.

³See §3.1 and §A.1 for our prompting implementation.

source	consistent	inconsistent
T5-11B	233,815	39,423
T5-3B	229,097	45,662
T5-large	195,681	81,986
T5-base	161,177	118,480
T5-small	88,129	190,012
total	907,899	475,563

Table 1: Our generated dataset statistics.

are then used as a synthetic example for training a factual consistency classifier. Since we leverage LLMs for labeling, our approach is likely to benefit from the ongoing progress in LLMs quality. Furthermore, previous approaches often rely on language-specific components (e.g., Information Extraction), which limits their applicability in multiple languages. Since recent LLMs are pretrained on multilingual data, our method can be easily applied to non-English languages, as we show in §5.

3 Experimental Setup

We apply TrueTeacher to generate a large-scale synthetic dataset for factual consistency evaluation in summarization and experiment with the resulted dataset to evaluate the effectiveness and the practical usefulness of our method (§4).

3.1 Synthetic Data Generation Process

To apply TrueTeacher, we instantiate the summarization datasets T , the pre-trained LMs and documents corpus D as follows. We use XSum (Narayan et al., 2018) as T , T5 (Raffel et al., 2020) pre-trained models as $LMs = \{T5\text{-small}, T5\text{-base}, T5\text{-large}, T5\text{-3B}, T5\text{-11B}\}$, and the documents in the CNN/DailyMail dataset (Yu et al., 2021) as D .

We employ FLAN-PaLM 540B (Chung et al., 2022) as our LLM-based teacher. FLAN-based models are trained to follow instructions in natural language, and FLAN-PaLM was fine-tuned on the closely-related NLI task.⁴ Therefore, we expect it to generalize well to factual consistency evaluation.⁵ We use zero-shot prompting for simplicity, and since applying few-shot or chain-of-thought prompting did not improve performance in early experiments. Additional implementation details of our FLAN-PaLM usage are provided in the Appendix (§A.1).

⁴https://github.com/google-research/FLAN/blob/main/flan/task_splits.py#L109

⁵We validate this in §4.1.

Applying TrueTeacher in this setup resulted in 1.4M synthetic examples (see statistics in Table 1). As expected, larger models output more consistent examples. To foster further research we make this dataset publicly available. We leverage this dataset to train a student model, that predicts a factual consistency label for a document-summary pair. We provide the implementation details for our trained models in §A.2.

3.2 Evaluation

To compare between consistency evaluation models, we use the TRUE benchmark (Honovich et al., 2022), focusing on its summarization subset: **MNBM** (Maynez et al., 2020), **FRANK** (Pagnoni et al., 2021), **SummEval** (Fabbri et al., 2020), **QAGS-X** and **QAGS-C** (Wang et al., 2020). For additional details about these datasets, we refer the reader to Honovich et al. (2022). Following Honovich et al., we use ROC-AUC in a binary classification setting as our evaluation metric.

3.3 Baselines

As baseline consistency evaluation models that did not train on synthetically generated data, we take the top performing methods from the TRUE study: **QuestEval** (Scialom et al., 2021), **Q²** (Honovich et al., 2021), **SUMMAC_{ZS}** (Laban et al., 2022), **T5-11B fine tuned on ANLI** (best performing single-model from the study), as well as the **Ensemble** of the 3 top performing models. We refer the reader to Honovich et al. (2022) for a detailed description of each method.

We compare TrueTeacher to the following synthetic data generation approaches:

DocNLI (Yin et al., 2021). Data is generated by reformatting NLI, Question Answering and Summarization datasets, including the CNN/DM corpus. The summarization-based positive examples are created using all the gold summaries concatenated as the hypothesis. The negative examples are generated by applying word/entity/sentence replacements to the positive hypothesis.

FactCC (Kryscinski et al., 2020). The premise is based on documents from CNN/DM. The positive hypotheses are based on randomly sampled sentences from the premise, some of them are paraphrased using back-translation and all of them are injected with noise by duplicating or removing random tokens. The negative hypotheses are then ob-

tained by rule-based transformations, such as sentence negation and entity/pronoun/number swaps.

Falsesum (Utama et al., 2022). The positive examples are based on documents and gold summaries from CNN/DM. For the negative examples the OpenIE (Banko et al., 2007) framework is utilized to identify predicates and arguments in the document and the gold summary. Randomly selected predicates and arguments from the gold summary are then masked and infilled using predicates and arguments from the document, or by "hallucinating" new content. For this purpose a dedicated infilling model is trained.

FactEdit (Balachandran et al., 2022). The positive examples are based on documents and gold summaries from CNN/DM. For the negative examples, an infilling model is trained based on sentences from the documents, by utilizing the OpenIE (Banko et al., 2007) framework to identify subjects, objects and relations and masking them. Each subject/object/relation phrase in the gold summary is then iteratively masked and infilled using the model’s lower order beam candidates.

4 Experiments and Analysis

We design our experiments to address the following research questions (RQs):

- **RQ1:** What is the performance of FLAN-PaLM 540B in factual consistency evaluation in summarization? Is it a good choice as a LLM teacher?
- **RQ2:** Can TrueTeacher facilitate training of a competitive model w.r.t. state-of-the-art models?
- **RQ3:** What is the quality of the data generated using TrueTeacher compared to existing synthetic data generation methods?

We address RQ1 and RQ2 in §4.1. To address RQ1, we evaluate FLAN-PaLM against competitive models for factual consistency evaluation. To address RQ2, we use our full dataset from §3.1 to train our best-performing model, and evaluate it in the exact same setting. Finally, RQ3 is addressed in §4.2, where we conduct a systematic study, comparing existing methods to TrueTeacher while controlling for factors such as the synthetic data size and the documents used for data synthesis.

4.1 Main Results on the TRUE Benchmark

We address RQ1 by evaluating FLAN-PaLM on the task and present the results in Table 2. FLAN-

	MNBM	QAGS-X	FRANK	SummEval	QAGS-C	Average
QuestEval (Scialom et al., 2021)	65.3	56.3	84.0	70.1	64.2	68.0
Q ² (Honovich et al., 2021)	68.7	70.9	87.8	78.8	83.5	77.9
SUMMAC _{ZS} (Laban et al., 2022)	71.3	78.1	89.1	81.7	80.9	80.2
T5-11B w. ANLI (Honovich et al., 2022)	77.9	83.8	82.1	80.5	89.4	82.7
Ensemble (Honovich et al., 2022)	76.6	85.8	91.2	82.9	87.7	84.8
FLAN-PaLM 540B	76.0	88.1	91.4	83.7	85.2	84.9
T5-11B w. ANLI + TrueTeacher full	78.1	89.4	93.6	88.5	89.4	87.8

Table 2: ROC-AUC results on the summarization subset of the TRUE benchmark (Honovich et al., 2022).

PaLM 540B achieves an impressive performance, with ROC-AUC of **84.9** compared to **82.7** of the best state-of-the-art single-model baseline, and performs on-par with an ensemble of the three best-performing models (Honovich et al., 2022). This demonstrates FLAN-PaLM’s capability for the task, and its potential as a teacher for smaller models.

To address RQ2, we fine-tune T5-11B (Raffel et al., 2020) over our full dataset (§3.1) mixed with ANLI (Nie et al., 2020). Table 2 shows that including TrueTeacher data in the training set, substantially improves the state-of-the-art single-model baseline from an average ROC-AUC of **82.7** to **87.8** (+5.1), while maintaining exactly the same model capacity. This strong result demonstrates the high effectiveness of TrueTeacher in a challenging setup. Notably, our model also outperforms the $\times 50$ times larger FLAN-PaLM that we used as the teacher ($84.9 \rightarrow 87.8$). This can be attributed to large-scale knowledge distillation on a specific task, which represents the "know-how" of the large model for that task, without the need to maintain performance on other tasks.

4.2 Re-evaluating Synthetic Data Generation Methods – A Study

Previous studies on synthetic data generation have used different experimental setups, making it difficult to compare their results. In this section, we design a systematic study to re-evaluate existing methods in a standardized setup. We first discuss our study design choices followed by the results.

Previous work demonstrated that synthetic data can improve NLI-based models. However, each work used a different model that often had relatively small capacity, while Honovich et al. (2022) recently showed that significant performance gains can be obtained by scaling to a T5-11B fine tuned on ANLI. We therefore use this **competitive baseline**, to which we add synthetic data from each method. For ablation purposes, we include variants

trained on synthetic data only, as well as with a smaller capacity T5-base variant.

To preform a fair comparison, we **restrict the number of examples** from each evaluated method to 100k, randomly sampled with balanced labels.

To evaluate **domain-shift robustness**, we further restrict the synthetic *training* examples to ones that were generated only based on CNN/DM documents,⁶ and then consider the XSum-based evaluation sets as out-of-domain.⁷

Table 3 presents the results of our study. We calculate three average scores: for in-domain test sets based on CNN/DM documents, for out-of-domain test sets based on XSum documents, and for the original datasets from TRUE.

In-Domain Results The majority of methods outperform the corresponding ANLI-based baseline, demonstrating the usefulness of synthetic data for the task. As expected, all the methods improve with larger models and a complementary effect is observed in the majority of the cases when the synthetic data is mixed with ANLI. The best results are obtained by mixing ANLI with either Falsesum or TrueTeacher data and using T5-11B, with a substantial improvement over the ANLI-only baseline (in-domain score increase from 81.1 to 87.9).

Out-of-domain Results While the majority of the methods perform well in-domain, their performance drops significantly on the out-of-domain test sets. Most of the baseline methods underperform the ANLI-only baseline using the same model capacity. For some methods, performance deteriorates dramatically, e.g., even though Falsesum

⁶Some methods are based exclusively on CNN/DM while others use additional datasets, more details in §3.3.

⁷SummEval and QAGS-C are based on documents from CNN/DM, MNBM and QAGS-X use documents from XSum, and FRANK has documents from both CNN/DM and XSum. We split FRANK to FRANK-C and FRANK-X which contain its CNN/DN based and XSum based subsets.

Training data	CNN/DM-based				XSUM-based			Average scores		
	QAGS-C	SummEval	FRANK-C	FRANK	FRANK-X	QAGS-X	MNBM	In-domain	Out-of-domain	TRUE
ANLI	83.4	74.2	85.6	90.7	93.2	88.0	73.9	81.1	85.0	82.0
T5-11B	FactEdit	87.8	77.0	77.2	83.7	76.0	69.4	80.7 (-0.4)	66.2 (-18.8)	74.2 (-7.8)
	FactEdit + ANLI	88.9	78.9	81.1	88.0	86.1	76.2	83.0 (+1.9)	74.0 (-11.0)	78.4 (-1.6)
	DocNLI	89.1	72.9	83.0	89.2	92.4	83.8	81.7 (+0.6)	81.1 (-3.9)	80.4 (-1.6)
	DocNLI + ANLI	87.8	72.0	81.9	88.2	93.7	84.2	80.6 (-0.5)	82.0 (-3.0)	80.0 (-2.0)
	FactCC	83.1	79.0	81.6	84.1	67.5	72.7	81.2 (+0.1)	65.1 (-19.9)	74.8 (-7.2)
	FactCC + ANLI	84.7	83.3	84.7	89.5	89.6	82.9	84.2 (+3.1)	81.3 (-3.7)	82.4 (+0.4)
	Falsesum	90.3	85.4	85.8	89.8	84.5	70.8	87.2 (+6.1)	69.7 (-15.3)	78.0 (-4.0)
	Falsesum + ANLI	90.7	85.8	87.0	91.6	90.5	75.2	87.8 (+6.7)	75.4 (-9.6)	80.8 (-1.2)
	TrueTeacher	84.9	85.0	88.8	93.6	94.4	86.5	86.2 (+5.1)	85.7 (+0.7)	85.2 (+3.2)
	TrueTeacher + ANLI	88.4	85.8	89.6	93.9	93.9	87.8	87.9 (+6.8)	86.0 (+1.0)	86.4 (+6.4)
T5-base	ANLI	74.9	63.7	73.1	81.3	80.6	77.2	70.6	78.3	74.8
	FactEdit	61.4	59.4	59.4	73.6	51.9	48.0	60.1 (-10.5)	52.8 (-25.5)	60.2 (-14.6)
	FactEdit + ANLI	68.7	60.0	62.2	78.5	73.6	72.2	63.6 (-7.0)	73.8 (-4.5)	71.0 (-3.8)
	DocNLI	71.4	66.5	66.7	77.9	81.0	75.2	68.2 (-2.4)	75.9 (-2.4)	72.5 (-2.3)
	DocNLI + ANLI	75.2	66.7	74.4	84.9	83.3	78.7	72.1 (+1.5)	78.9 (+0.6)	76.1 (+1.3)
	FactCC	74.0	72.7	78.7	83.2	71.9	71.0	75.3 (+4.7)	68.5 (-9.8)	72.7 (-2.1)
	FactCC + ANLI	72.8	73.2	78.8	83.2	66.8	71.5	74.9 (+4.3)	67.2 (-11.1)	72.8 (-2.0)
	Falsesum	80.9	74.2	82.0	86.4	71.6	65.0	79.0 (+8.4)	63.2 (-15.1)	71.9 (-2.9)
	Falsesum + ANLI	82.9	73.4	83.3	86.5	72.6	66.0	79.9 (+9.3)	65.8 (-12.5)	73.5 (-1.3)
	TrueTeacher	77.3	73.6	79.1	88.0	82.6	79.9	76.7 (+6.1)	80.3 (+2.0)	79.4 (+4.6)
	TrueTeacher + ANLI	81.9	78.0	81.4	89.3	86.4	81.9	80.4 (+9.8)	82.3 (+4.0)	81.9 (+7.1)

Table 3: ROC-AUC results on TRUE comparing different synthetic data generation methods. For each model size, average scores are compared to the corresponding ANLI-only baseline (difference is listed in parentheses).

presents major in-domain gains it significantly underperforms the ANLI-only baseline. This suggests that some methods may overfit to documents from the distribution used to generate the synthetic data. Based on this finding, we suggest that future research should prioritize out-of-domain evaluation. Interestingly, even though TrueTeacher’s relative improvement is smaller compared to the in-domain setup, it is still the only method with better out-of-domain score than the ANLI-only baseline. This suggests that TrueTeacher is robust to domain shift, which may be due to the use of model-generated summaries that increase the variability of the resulting synthetic data.

Overall Results on TRUE Due to the poor out-of-domain performance of the existing methods, TrueTeacher is the only method that consistently outperforms the ANLI only baseline on the TRUE benchmark. Notably, TrueTeacher + ANLI with T5-base (81.9) performs on par with the ANLI-only baseline using T5-11B (82.0). Additionally, the TrueTeacher-based variant using T5-11B (85.2) already performs on-par with FLAN-PaLM 540B (84.9), even though we used only 100k synthetic examples in this experiment, and did not use ANLI data. When comparing TrueTeacher + ANLI with T5-11B and 100k examples (Table 3) to the equivalent variant using the full dataset (Table 2), we observe a performance increase (86.4 \rightarrow 87.8), which demonstrates TrueTeacher’s scalability. We conclude that TrueTeacher yields high quality data

and generalizes well for new domains, which we attribute to the usage of model-generated summaries.

4.3 Qualitative Analysis

Figure 3 presents a case study, with a randomly sampled (not cherry-picked) *negative* example from all the evaluated methods, based on the same document. **FactEdit** used the second gold-summary and replaced “to flooding call” with “rescue”, introducing a grammatical error rather than a clear factuality error, demonstrating that using lower-beam completions as proxy for factuality errors can be problematic. **DocNLI** uses all the gold summaries concatenated. While replacing “morning” with “night” introduces a factual inconsistency, three other edits fail to introduce factual inconsistencies, which demonstrates the limitations of using simple word/entity replacements. **FactCC** uses random sentences from the article, in this case the first one, and introduces a factual error by an entity swap from “firetruck” to “fire engine”. The paraphrase highlighted in green helps increases the abstractiveness, but the paraphrase highlighted in orange introduces a grammatical error that is less likely to be made by a strong summarization model. The noise injection used by FactCC (duplicating or removing random tokens) is colored in red, but its usefulness is questionable. **Falsesum** uses the first gold summary. The representation of the summary as a set of predicates and arguments allows larger edits, e.g. the removal of “Tuesday morning”. Falsesum replaces the “sinkhole” argument with

CNN/DailyMail ID: 372f7e02e5bb17bac3a1b2260c6ac78414f97ee3

Article: LOS ANGELES, California (CNN) -- Los Angeles firefighters and city crews worked for several hours Tuesday to rescue one of their own: a 22-ton firetruck that was nearly swallowed by a water-logged sinkhole. Two firefighters crawled out of the truck's windows after it sank Tuesday morning. No one was injured. The incident happened after four firefighters took the truck to the San Fernando Valley neighborhood of Valley Village, where flooding had been reported... ..

Gold Summaries:

1. Los Angeles firetruck nearly swallowed by sinkhole Tuesday morning.
2. Firefighters in truck were responding to flooding call when incident happened.
3. Two firefighters escaped truck through windows; no injuries reported.

FactEdit	Firefighters in truck were responding rescue when incident happened .
DocNLI	Los Angeles firetruck nearly destroyed by sinkhole Tuesday night . Firefighters in truck were responding to emergency call when it happened . Two firefighters escaped truck through windows ; no injuries reported .
FactCC	LOS LOS ANGELES, California ((CNN) - Los Angeles firefighters and crews worked Two on Tuesday to rescue one of their own: a 22-ton fire engine nearly swallowed by a sinkhole filled with waterwater.
Falsesum	Los Angeles firetruck nearly swallowed by water.
TrueTeacher	A firefighter has rescued a truck that sank in Los Angeles, causing extensive flooding.

Figure 3: A case study comparing factually inconsistent summaries of the same document generated using different methods. Content replacements are highlighted using the same color for the original and the replaced text. Added content is highlighted in red.

"water", failing to introduce a factual inconsistency, since the sinkhole is referred to as "water-logged sinkhole" in the article. Finally, **TrueTeacher** uses an abstractive summary generated by a real summarization model. It introduces a nuanced factual inconsistency by replacing "Los Angeles firefighters" with *A firefighter* and also by hallucinating new content (the text in bold red font). It shows factual errors produced by a real summarization model that are more likely to occur. This case study further illustrates the challenges of perturbing texts to introduce factual inconsistencies and re-iterates the importance in using model-generated summaries.

4.4 Abtractiveness Analysis

Advances in large scale pretraining (Devlin et al., 2019; Lewis et al., 2020) and the availability of relevant datasets (Narayan et al., 2018), enabled rapid progress in *abstractive* summarization, which better imitates the way humans summarize (Koh et al., 2023) and is also preferred by humans (Goyal et al., 2022). However, abstractive summarization presents a great challenge for generative models, with up to 30% of the generated summaries containing factual inconsistencies (Cao et al., 2018; Kryscinski et al., 2019). This motivates us to focus on generating *abstractive* synthetic summaries.

	coverage ↓	density ↓	combined ↓
FactEdit	0.859	2.923	2.671
DocNLI	0.845	15.656	15.203
FactCC	0.928	8.155	7.928
Falsesum	0.882	2.977	2.756
TrueTeacher	0.856	2.406	2.152

Table 4: Average abtractiveness scores (lower is better), measured on a random sample of 5k examples.

We compare the abtractiveness degree of different methods using the extractive fragment *coverage*⁸ and *density*⁹ measures defined by Grusky et al. (2018). Following Utama et al. (2022) we multiply these measures to obtain a *combined* score.

Table 4 presents the average abtractiveness scores, and we also included a density plot of the combined scores in the Appendix (Figure 4). We observe higher abtractiveness for model-based methods (FactEdit, Falsesum and TrueTeacher), suggesting that rule-based methods might be less relevant with the recent shift towards abstractive summarization. TrueTeacher produces the most abstractive summaries with lowest combined score.

5 Multi-Lingual Data Generation for Factual Consistency Evaluation

Leveraging a multilingual LLM as the teacher allows us to apply TrueTeacher to multiple languages in a straightforward manner. This is in contrast to approaches that rely on NLP components that are only available for high-resource languages, e.g., information extraction (Utama et al., 2022; Balachandran et al., 2022). In this section, we explore the usefulness of TrueTeacher for multilingual factual consistency evaluation. We first generate multilingual synthetic data using TrueTeacher. This time we train a single summarization model by fine tuning mT5-XXL (Xue et al., 2021) on XLSum (Hasan et al., 2021) and use it to summarize documents from WikiLingua (Ladhak et al., 2020), which we then label for consistency with FLAN-PaLM. For the purposes of this experiment we focus on a subset of WikiLingua documents in 4 languages: English (en), French (fe), Spanish (es) and German (de), since they are the most prevalent in PaLM’s training data.¹⁰ After generating the dataset for

⁸Measures the percentage of words in the summary that are part of an extractive fragment in the document.

⁹Measures the average extractive fragment length to which each word in the summary belongs.

¹⁰See Table 29 in Chowdhery et al. (2022).

Training data	# improved languages	avg. ROC-AUC per lang.	per ex.
ANLI+XNLI	-	73.3	71.6
+TrueTeacher en	32 / 45	75.7	73.8
+TrueTeacher en,fe,es,ge	35 / 45	77.2	75.3

Table 5: Multilingual results on mFACE test set.

these 4 languages, we sample 100k examples, by randomly sampling 25k in each language with balanced labels (as illustrated in Table 6 in the Appendix). For ablation purposes, we also create an English-only variant, by randomly sampling 100k English examples with balanced labels.¹¹

We then use the resulted data to train models for factual consistency evaluation in multiple languages and evaluate them on the test set from mFace (Aharoni et al., 2022), containing 3150 examples in 45 languages. As a strong baseline we follow Aharoni et al. and fine-tune mT5-XXL (Xue et al., 2021) on the ANLI (Nie et al., 2020) and XNLI (Conneau et al., 2018) datasets. We then assess whether adding synthetic data generated with TrueTeacher to the training set can improve this model, and we compare the multi-lingual data to the English only ablated variant.

Table 5 presents the results overview, full results in all 45 languages are available in Table 7 (Appendix). Adding English-only summarization-based synthetic data, already improves results on 32 out of 45 languages and increases the avg. ROC-AUC from 71.6 to 73.8. Yet, using the same amount of multi-lingual examples improved the performance even more, with avg. ROC AUC of 75.3. This demonstrates the added value in generating multi-lingual synthetic examples using TrueTeacher, laying the ground for future work.

6 Related Work

Previous work proposed methods for generating synthetic training data for factual consistency evaluation, by perturbing gold summaries (Yin et al., 2021; Kryscinski et al., 2020; Balachandran et al., 2022; Utama et al., 2022).¹² A key advantage of TrueTeacher is the ability to leverage real model-generated summaries. While our research was in progress, another closely related work proposed to leverage model-generated summaries, and to label them by aggregating scores from multiple existing

metrics (Wu et al., 2022).¹³ Our work proposes a significantly simpler method that is also multilingual by nature, and provides valuable insights into the robustness of existing methods through our systematic study (§4.2).

Another line of work for adapting NLI-based models for summarization, focuses on better processing of long texts, splitting the documents into sentences to create shorter premise-hypothesis pairs (Laban et al., 2022; Schuster et al., 2022).

With the increasing interest in LLMs, recent work attempts to assess their capability for evaluating generation tasks (Kocmi and Federmann, 2023; Wang et al., 2023; Liu et al., 2023). Luo et al. (2023) focused specifically on the task of factual consistency evaluation in summarization, evaluating ChatGPT’s (OpenAI, 2022) performance on the task. Yet, Aiyappa et al. (2023) argued that the usage of ChatGPT is problematic since its "closed" nature makes it impossible to ensure there was no data leakage (training-test contamination).¹⁴ Previous work also leveraged LLMs for data annotation (Wang et al., 2021; Ding et al., 2022), as well as synthetic data generation (Agrawal et al., 2022; Liu et al., 2022; Bitton et al., 2023). As far as we aware our work is the first to leverage LLMs for data generation for factual consistency.

7 Conclusion

We introduced TrueTeacher, a simple and highly effective method for generating synthetic data for factual consistency evaluation. Instead of perturbation of human-written summaries like done in previous work, TrueTeacher leverages realistic model-generated summaries, which are annotated using a Large Language Model.

Adding our data to the training set substantially improves the state-of-the-art model on the TRUE benchmark, while maintaining similar model capacity. Our experiments demonstrated that our approach generalizes well to multilingual scenarios and new domains. In a systematic study of existing synthetic data generation methods, we uncover their limitations and emphasize the necessity of out-of-domain evaluation. To foster further research, we release a large-scale synthetic dataset containing 1.4M examples generated using our method.

¹³As no data or code were published, we could not compare their method to ours.

¹⁴While FLAN’s instruction fine-tuning data is detailed here: <https://github.com/google-research/FLAN/blob/main/flan/tasks.py>

¹¹Also based on WikiLingua, generated with the same process like the 25k English subset of our multilingual dataset.

¹²We provide extensive review of these methods in §3.3.

References

- Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2022. [Qameleon: Multilingual QA with only 5 examples](#). *CoRR*, abs/2211.08264.
- Roei Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2022. [mface: Multilingual summarization with factual consistency evaluation](#). *CoRR*, abs/2212.10622.
- Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yongyeol Ahn. 2023. [Can we trust the evaluation on chatgpt?](#) *CoRR*, abs/2303.12767.
- Vidhisha Balachandran, Hannaneh Hajishirzi, William W. Cohen, and Yulia Tsvetkov. 2022. [Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9818–9830. Association for Computational Linguistics.
- Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. [Open information extraction from the web](#). In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 2670–2676.
- Yonatan Bitton, Shlomi Cohen-Ganor, Ido Hakimi, Yoav Lewenberg, Roei Aharoni, and Enav Weinreb. 2023. [q2d: Turning questions into dialogs to teach models how to search](#).
- Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. [Faithful to the original: Fact aware neural abstractive summarization](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#). *CoRR*, abs/2204.02311.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#). *CoRR*, abs/2210.11416.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq R. Joty, and Boyang Li. 2022. [Is GPT-3 a good data annotator?](#) *CoRR*, abs/2212.10450.
- 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. 2019a. [Ranking generated summaries by correctness: An interesting but challenging application for natural language inference](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2214–2220. Association for Computational Linguistics.
- Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019b.

- 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.
- Ben Goodrich, Vinay Rao, Mohammad Saleh, and Peter J. Liu. 2019. [Assessing the factual accuracy of generated text](#). *CoRR*, abs/1905.13322.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. [News summarization and evaluation in the era of GPT-3](#). *CoRR*, abs/2209.12356.
- Max Grusky, Mor Naaman, and Yoav Artzi. 2018. [Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 708–719. Association for Computational Linguistics.
- Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Samin Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. [Xl-sum: Large-scale multilingual abstractive summarization for 44 languages](#). In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4693–4703. Association for Computational Linguistics.
- Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. [TRUE: re-evaluating factual consistency evaluation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3905–3920. Association for Computational Linguistics.
- Or Honovich, Leshem Choshen, Roei Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. [\\$q^2\\$: 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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7856–7870. Association for Computational Linguistics.
- Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.
- Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. [Scitail: A textual entailment dataset from science question answering](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 5189–5197. AAAI Press.
- Tom Kocmi and Christian Federmann. 2023. [Large language models are state-of-the-art evaluators of translation quality](#). *CoRR*, abs/2302.14520.
- Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2023. [An empirical survey on long document summarization: Datasets, models, and metrics](#). *ACM Comput. Surv.*, 55(8):154:1–154:35.
- 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 2019, Hong Kong, China, November 3-7, 2019*, pages 540–551. 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 2020, Online, November 16-20, 2020*, pages 9332–9346. Association for Computational Linguistics.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. [Summac: Re-visiting nli-based models for inconsistency detection in summarization](#). *Trans. Assoc. Comput. Linguistics*, 10:163–177.
- Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen R. McKeown. 2020. [Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization](#). *CoRR*, abs/2010.03093.
- 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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.
- Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. [WANLI: worker and AI collaboration for natural language inference dataset creation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6826–6847. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval:](#)

- NLG evaluation using GPT-4 with better human alignment. *CoRR*, abs/2303.16634.
- Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. [Chatgpt as a factual inconsistency evaluator for abstractive text summarization](#). *CoRR*, abs/2303.15621.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.
- Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. [Looking beyond sentence-level natural language inference for question answering and text summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1322–1336. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807. 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, ACL 2020, Online, July 5-10, 2020*, pages 4885–4901. Association for Computational Linguistics.
- OpenAI. 2022. [Chatgpt, https://openai.com/blog/chatgpt/](https://openai.com/blog/chatgpt/).
- 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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4812–4829. 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](#). *J. Mach. Learn. Res.*, 21:140:1–140:67.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2021. [Measuring attribution in natural language generation models](#). *CoRR*, abs/2112.12870.
- Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. [Stretching sentence-pair NLI models to reason over long documents and clusters](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 394–412. 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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6594–6604. Association for Computational Linguistics.
- Prasetya Utama, Joshua Bambrick, Nafise Sadat Moosavi, and Iryna Gurevych. 2022. [Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2763–2776. 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, ACL 2020, Online, July 5-10, 2020*, pages 5008–5020. Association for Computational Linguistics.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. [Is chatgpt a good NLG evaluator? A preliminary study](#). *CoRR*, abs/2303.04048.
- Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. [Want to reduce labeling cost? GPT-3 can help](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4195–4205. Association for Computational Linguistics.
- Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Sujian Li, and Yajuan Lv. 2022. [Wecheck: Strong factual consistency checker via weakly supervised learning](#). *CoRR*, abs/2212.10057.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mt5: A massively](#)

[multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Wenpeng Yin, Dragomir R. 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, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4913–4922. Association for Computational Linguistics.

Dian Yu, Kai Sun, Dong Yu, and Claire Cardie. 2021. [Self-teaching machines to read and comprehend with large-scale multi-subject question-answering data](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 56–68. Association for Computational Linguistics.

A Appendix

A.1 Prompting FLAN-PaLM

To apply FLAN-PaLM for factual consistency classification, we use the following prompt in a zero-shot format:

Premise: <document> Hypothesis:
<summary> Can the hypothesis be inferred from the premise? Answer using "Yes" or "No" only.

The vast majority of Flan-PaLM responses were either Yes or No, and a tiny fraction of the responses were It's impossible to say.

During the labeling phase with FLAN-PaLM we let the model generate the output and label as positive if the output is Yes and negative if the output is No. We discard It's impossible to say examples. In order to measure ROC-AUC in a binary classification setting, we compute the model's probability of generating Yes and use it as the example level factual consistency score.

A.2 Fine tuning T5

We fine tune our T5 models for factual consistency evaluation using the following input format:

premise: <document> hypothesis:
<summary>

The model is trained to predict "1" if the summary is factually consistent and "0" otherwise. We use a learning rate of 10^{-4} and a batch size of 32. During training, we use a maximum input length of 512 tokens and truncate the premise if needed.¹⁵ During inference we use a maximum input length of 2048 tokens. We train for a maximum of 20 epochs and choose the checkpoint with the best ROC-AUC score on a development set.¹⁶ For development set we use the FactCC dataset (Kryscinski et al., 2020) with 1,431 examples containing summaries of documents from CNN/DailyMail, manually annotated for factual correctness.¹⁷ In our study we make sure to use the same training regime for all baselines.

The ANLI-only results in Tab. 3 are from our experiments, while in Tab. 2 we use the results from Honovich et al. (2022).

For the summarization models we fine tune the corresponding T5 models on the XSum training set

¹⁵In early experiments we saw that training with longer maximum input length resulted with comparable performance.

¹⁶We evaluate a checkpoint every 1k steps.

¹⁷Following (Utama et al., 2022), we merge the dev and test sets.

Language	ISO 639-1	consistent	inconsistent
English	en	12,500	12,500
Spanish	es	12,500	12,500
French	fr	12,500	12,500
German	de	12,500	12,500
total		50,000	50,000

Table 6: Our multilingual dataset statistics.

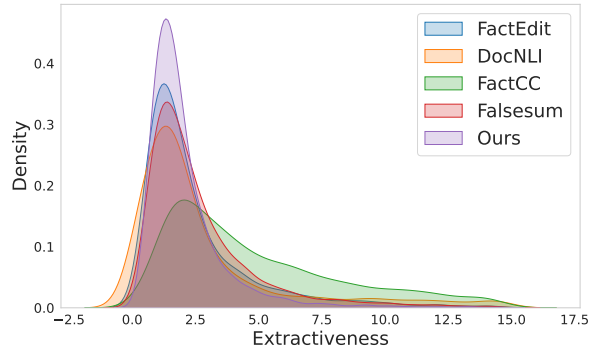


Figure 4: Visualization of the density of the combined abstractivness score. The plot is actually measuring the extractiveness degree, so lower x-values mean higher abstractivness.

(Narayan et al., 2018) in a similar fashion and use the ROUGE score on the XSum development set as a stopping criteria.

A.3 Additional details about our generated dataset

As mentioned in §3.1, we create the dataset based on documents from CNN/DailyMail (Yu et al., 2021). We do not use the gold summaries, and we only use examples from the training set.

In our experiments with the full dataset (§4.1), we balance the labels by randomly sampling 475,563 positive examples (see Table 1).

A.4 Using the mFace dataset

In §5 we report results on the mFace dataset (Aharoni et al., 2022). Aharoni et al. performed large scale human evaluation of summaries of documents from the XLSum corpus (Hasan et al., 2021), produced by different summarization models. Each summary was rated for quality, attribution and informativeness. We use the attribution scores in our work. The attribution evaluation is based on the attribution definition provided in Rashkin et al. (2021), with the participants asked "Is all the information in the summary fully attributable to the article?". In our work we use the average attribu-

	ANLI+XNLI	+100K en	+100K en/es/de/fe
amharic	63.1	67.2	68.6
arabic	87.8	89.0	87.7
azerbaijani	59.6	68.6	65.5
bengali	90.4	94.3	98.5
burmese	59.0	64.5	57.9
chinesesimp.	87.6	86.4	89.9
chinese trad.	82.5	82.6	83.2
english	80.2	74.7	80.0
french	91.9	94.1	97.1
gujarati	50.8	52.0	51.5
hausa	69.5	67.7	73.7
hindi	72.2	79.9	86.5
igbo	62.2	62.8	75.7
indonesian	77.6	84.1	85.8
japanese	97.7	98.9	99.6
kirundi	83.5	89.3	90.4
korean	87.3	82.3	89.9
kyrgyz	70.1	77.4	79.0
marathi	75.2	78.7	73.6
nepali	55.2	59.1	57.2
oromo	81.2	83.7	83.3
pashto	56.4	68.2	67.7
persian	43.5	42.3	45.8
pidgin	70.0	81.4	77.1
portuguese	79.6	79.5	79.0
punjabi	77.7	81.5	78.2
russian	88.8	85.1	81.2
scottish gaelic	59.0	58.8	63.1
serbian cyrillic	84.2	79.3	85.5
serbian latin	39.7	42.2	43.6
sinhala	72.9	74.9	76.1
somali	85.1	88.6	86.6
spanish	80.7	85.9	89.1
swahili	88.1	89.2	92.2
tamil	63.9	69.8	66.0
telugu	55.9	62.3	60.4
thai	78.8	83.8	86.8
tigrinya	79.9	82.9	86.1
turkish	87.0	86.6	86.6
ukrainian	55.5	67.0	65.9
urdu	69.0	63.8	75.3
uzbek	54.6	59.3	58.8
vietnamese	89.8	84.4	88.1
welsh	83.0	83.4	83.9
yoruba	69.0	69.0	77.2
# wins	5	15	25
# > ANLI+XNLI	-	32	35
Per lang. avg.	73.3	75.7	77.2
Per example avg.	71.6	73.8	75.3

Table 7: ROC-AUC results on the mFace test set (Aharoni et al., 2022).

tion score (between 0 to 1) and treat summaries as factually consistent if the score is larger than 0.5. We focus on the test split of XLSum containing 3150 examples in 45 languages (i.e., 70 examples in each language). In §5 we refer to Table 5 with the results overview, and we provide the full results for all languages in Table 7.