Multilingual Summarization with Factual Consistency Evaluation

Roee Aharoni[*]
Google Research
roeeaharoni@google.com

Shashi Narayan[*]
Google DeepMind
shashinarayan@google.com

Joshua Maynez
Google DeepMind
joshuahm@google.com

Jonathan Herzig
Google Research
jherzig@google.com

Elizabeth Clark
Google DeepMind
eclark@google.com

Mirella Lapata
Google DeepMind
lapata@google.com

Abstract

Abstractive summarization has enjoyed renewed interest in recent years, thanks to pretrained language models and the availability of large-scale datasets. Despite promising results, current models still suffer from generating factually inconsistent summaries, reducing their utility for real-world application. Several recent efforts attempt to address this by devising models that automatically detect factual inconsistencies in machine generated summaries. However, they focus exclusively on English, a language with abundant resources. In this work, we leverage factual consistency evaluation models to improve *multilingual* summarization. We explore two intuitive approaches to mitigate hallucinations based on the signal provided by a multilingual NLI model, namely data filtering and controlled generation. Experimental results in the 45 languages from the XLSum dataset show gains over strong baselines in both automatic and human evaluation. We release models and human judgements of summaries to foster progress towards more factually consistent multilingual summarization.[1]

1 Introduction

The past few years have witnessed a huge leap forward in abstractive summarization thanks to large-scale pretraining (Devlin et al., 2019; Lewis et al., 2020) and the availability of benchmark datasets. A well-known issue limiting the wider adoption of abstractive summarization models is their tendency to generate factually inconsistent summaries, a.k.a "hallucinations" (Maynez et al., 2020; Zhao et al., 2020, *inter alia*). A recently popular line of work explores how to best detect hallucinations in machine generated text, thereby enabling the automatic identification of factually inconsistent summaries (Eyal et al., 2019; Falke et al., 2019; Kryscinski et al., 2020; Wang et al., 2020; Goyal and Durrett, 2021; Scialom et al., 2021; Honovich et al., 2022; Tang et al., 2022, *inter alia*).

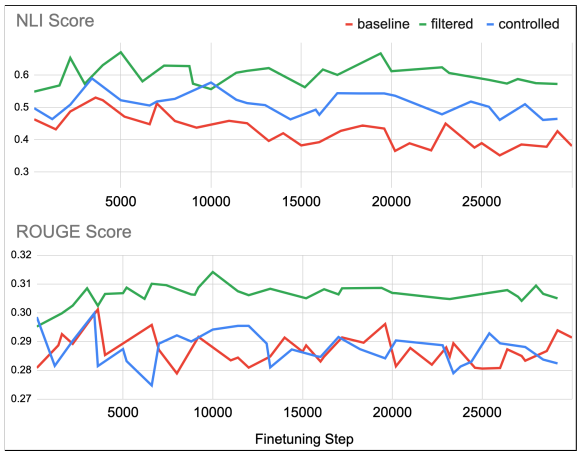


Figure 1: NLI and ROUGE scores for different models on the Arabic development set of XLSUM during finetuning. Using a multilingual entailment model during training (via data filtering or controlled generation) improves summary quality over a baseline model trained without using the entailment signal.

While such approaches may prove useful for automatic evaluation, it remains unclear how to best leverage them for *improving* summarization models *in multiple languages*. While focusing exclusively on English, previous work suggests different techniques to this effect such as discarding "noisy" training examples (Gehrmann et al., 2021), contrastive learning paradigms (Nan et al., 2021b), controlled generation and planning (Narayan et al., 2021; Rashkin et al., 2021b), or reinforcement-learning approaches that use the evaluation model score as a reward function (Gunasekara et al., 2021). Despite promising results, no method has emerged as a clear winner in English, let alone across languages with varying amounts of data and resources.

In this work, we leverage factual consistency evaluation models to improve summarization systems in multiple languages. Specifically, we employ Textual Entailment models (a.k.a. Natural

[*] Equal contribution.
[1] https://github.com/google-research/google-research/tree/master/mface

Findings of the Association for Computational Linguistics: ACL 2023, pages 3562–3591
July 9-14, 2023

Language Inference; Dagan et al. 2005; Bowman et al. 2015) in order to determine whether a summary is factually consistent (Maynez et al., 2020; Laban et al., 2021). We opportunistically opt for NLI given the availability of multilingual benchmarks for model training (Conneau et al., 2018; Nie et al., 2020). Approaches based on question generation and answering have been also shown to work well for factuality evaluation (i.e. Scialom et al., 2021; Honovich et al., 2021; Deutsch et al., 2021), however, they are not easily portable due to the scarcity of respective resources in languages other than English.

We first analyze the quality of the training data for summarization models using a strong multilingual NLI model (as evaluated on the XNLI dataset, Conneau et al.; 2018). In particular, we train our multilingual NLI model, following the guidelines from the TRUE survey (Honovich et al., 2022) for the assessment of factual consistency. Focusing on the XLSum[2] multilingual summarization dataset (Hasan et al., 2021), we find that for some languages up to 70% of training examples are not factually consistent according to the NLI model, while such examples are commonly used for training. We use the NLI signal to improve the quality of the generated summaries in two ways: (1) *data filtering*, where we only train on examples whose summaries are predicted to be entailed by the input, and (2) *controlled generation*, where we also leverage "negative" training examples by conditioning the summarization model on the NLI signal. We evaluate the proposed approaches using both automatic and human evaluation in 45 languages, and observe significant gains in the faithfulness of the generated summaries over strong baselines. Finally, we show that the human judgments we collected in all languages are useful for training automatic metrics to assess the quality, factual consistency and informativeness of generated summaries.

To summarize, the contributions of this work are three-fold: (1) we analyze the quality of the XLSum dataset (Hasan et al., 2021) using strong multilingual NLI models and reveal severe issues with faithfulness in the training data across languages; (2) we explore methods for improving downstream summarization models trained on this data using a multilingual NLI signal, and show large gains in both automatic and human evaluation; and (3) using the data from our large-scale human evaluation study, we learn metrics for automatically evaluating summaries in multiple languages along the dimensions of Quality, Factual Consistency, and Informativeness.[3] To the best of our knowledge, our work is the first to examine the faithfulness of summarization systems in multilingual settings, and we hope it will encourage the development of better metrics and models in multilingual text generation.

2 Related Work

Despite significant improvements in recent years (Liu and Lapata, 2019; Lewis et al., 2020; Roberts et al., 2020), abstractive summarization models are still prone to "hallucination", i.e., the inclusion of factual errors in the generated summaries (Song et al., 2018; Maynez et al., 2020; Kryscinski et al., 2020; Gabriel et al., 2021).

A plethora of approaches have been proposed for the automatic detection of factual inconsistencies in machine generated text (see Honovich et al. 2022 and Tang et al. 2022 for overviews) with varying degrees of success. There is growing consensus that techniques based on textual entailment (Maynez et al., 2020; Goyal et al., 2021; Goyal and Durrett, 2021) and question generation and answering models (Durmus et al., 2020; Wang et al., 2020; Deutsch et al., 2021; Fabbri et al., 2021; Scialom et al., 2021; Honovich et al., 2021) achieve strong performance across tasks and datasets (Laban et al., 2021; Honovich et al., 2022). Another line of work uses synthetically generated data to train models for evaluating factual consistency (Kryscinski et al., 2020; Zhao et al., 2020; Goyal and Durrett, 2020).

Aside from assessing system output, several studies have proposed novel model architectures which enforce factuality during training or inference. These include extracting facts from the source and incorporating them as additional input to the model (Cao et al., 2018; Aralikatte et al., 2021; Zhu et al., 2021), planning using entity chains and avoiding entities that are not in the input (Narayan et al., 2021, 2022), using reinforcement learning to optimize model training with factual correctness as a reward (Zhang et al., 2020; Arumae and Liu, 2019; Pasunuru and Bansal, 2018; Nan et al., 2021b), reranking candidate summaries

[2]Available for non-commercial use, see: http://github.com/csebuetnlp/xl-sum#license

[3]We will release our NLI models, summarization models, and human judgments of summaries in multiple languages to foster future work on this task.

within a beam using entailment predictions (Falke et al., 2019) or quantity verification scores (Zhao et al., 2020), using contrastive learning (Cao and Wang, 2021; Wan and Bansal, 2022), modifying the training objective to only maximize the likelihood of factual words (Goyal and Durrett, 2021), incorporating factuality into the pretraining objective of models tailored to text summarization tasks (Wan and Bansal, 2022), and adaptively removing examples with high log loss (Kang and Hashimoto, 2020). Other work simply removes noisy training samples (Nan et al., 2021a; Goyal and Durrett, 2021) in the hope that factuality will improve by training on better examples.

Despite promising results, it is unclear whether previous techniques transfer to languages beyond English. Our own work aims to improve the factuality of abstractive summarization *across languages*. Leveraging recent progress on multilingual pretrained models (Xue et al., 2021), we show that entailment-based metrics can be trained to detect factually-inconsistent summaries in multiple languages, and that this signal can be leveraged to improve summarization systems in those languages.

3 Multilingual Factual Consistency Evaluation

We cast factual consistency evaluation as a Natural Language Inference (NLI) task. The input forms the premise, the summary forms the hypothesis (Maynez et al., 2020; Laban et al., 2021; Honovich et al., 2022), and the NLI model is used to predict whether the summary is entailed by the input. More formally, given input document d and summary s, we define an NLI model $\mathcal{M}$ as a binary classifier, where $\mathcal{M}(d, s) \approx p(s \text{ is entailed by } d)$.

Recent studies (Honovich et al., 2022) on evaluating factual consistency in summarization and other related tasks in English have obtained promising results when finetuning large pretrained models on NLI datasets. Specifically, they finetune the T5 pretrained encoder-decoder models (Raffel et al., 2020) for binary classification where the entailment relation translates to a positive label and contradiction/neutral relations are merged to a negative label. Their model encodes the concatenation of the premise (document) and hypothesis (summary) and decodes a single token that represents the class label (entailment or no entailment).[4]

Since we are interested in evaluating factual consistency in multiple languages, we extend the modeling approach of Honovich et al. (2022) to a multilingual setting. As our pretrained model, we use mT5-XXL (Xue et al., 2021) which was trained on mC4, a dataset drawn from the public Common Crawl covering 101 languages. We finetuned mT5-XXL on the ANLI (Nie et al., 2020) and XNLI (Conneau et al., 2018) datasets. ANLI contains 162K English-only examples, while XNLI has 37K examples[5] in 15 languages. As mentioned above for the English case, the multilingual model is trained to generate a binary label when given the concatenation of a premise and hypothesis, where the positive label corresponds to an entailment relation, and the negative label stands for a neutral/contradiction relation. During inference, we score a premise and hypothesis input by measuring the output probability when force-decoding the positive label, resulting in a score between 0 (no entailment) and 1 (entailment).

We measured the quality of our multilingual NLI model by evaluating on the XNLI (Conneau et al., 2018) test set and the TRUE benchmark (Honovich et al., 2022). The latter is a standardized collection of datasets representing various tasks (summarization, dialog generation, paraphrasing and fact-checking) with manual annotations for factual consistency. On XNLI, our model yields an average accuracy of 90.0 over 15 languages in comparison to 87.8 reported in Xue et al. (2021).[6] We present results for individual languages in Appendix A. On the (English-only) TRUE benchmark, our model's average ROC AUC is 82.4 in comparison to 83.4 reported in Honovich et al. (2022) for their best performing English-only, T5-11B model (Raffel et al., 2020) trained on ANLI. While our model is trained on both ANLI (English) and XNLI (15 languages, detailed in Table 1), we assume it can generalize to additional languages (for which NLI data is not available) due to the nature of the pretrained model (mT5, trained on 101 languages).

4 Summarization Models

We next describe two summarization approaches which exploit the factual consistency evaluation

[4]Other work (Laban et al., 2021; Schuster et al., 2022) breaks documents into sentences before running NLI models, however, we refrained from doing so to avoid loss of context.

[5]We use the XNLI development set for training as it was manually curated, unlike the XNLI training set which was automatically translated.

[6]The numbers are not fully comparable as Xue et al. (2021) classify into three classes while we use binary classification.

Language	# Train	Ent%	Language	# Train	Ent%
Amharic	5,761	42.67	Pidgin	9,208	39.09
Arabic	37,519	45.19	Portuguese	57,402	40.19
Azerbaijani	6,478	34.15	Punjabi	8,215	28.29
Bengali	8,102	48.58	Russian	62,243	40.50
Burmese	4,569	32.48	Scottish (G)	1,313	42.42
Chinese (S)	37,362	50.81	Serbian (C)	7,275	31.08
Chinese (T)	37,373	50.59	Serbian (L)	7,276	30.61
English	306,522	60.00	Sinhala	3,249	36.53
French	8,697	36.96	Somali	5,962	38.49
Gujarati	9,119	29.60	Spanish	38,110	35.29
Hausa	6,418	39.42	Swahili	7,898	42.83
Hindi	70,778	43.68	Tamil	16,222	50.63
Igbo	4,183	35.64	Telugu	10,421	35.19
Indonesian	38,242	52.48	Thai	6,616	46.84
Japanese	7,113	68.96	Tigrinya	5,451	36.93
Kirundi	5,746	44.27	Turkish	27,176	44.31
Korean	4,407	48.92	Ukrainian	43,201	38.27
Kyrgyz	2,266	32.48	Urdu	67,665	42.17
Marathi	10,903	29.28	Uzbek	4,728	36.40
Nepali	5,808	51.58	Vietnamese	32,111	36.12
Oromo	6,063	40.76	Welsh	9,732	49.38
Pashto	14,353	44.83	Yoruba	6,350	37.15
Persian	47,251	49.19	Avg	24,670	41.37

Table 1: Statistics on XLSum training data: total number of examples per language, proporion of examples where the summary was entailed by the input (% Ent). Languages in XLSum *and* XNLI are underlined, for other lanuages NLI classification is zero-shot. Chinese (S/T) refers to simplified/traditional; Serbian (C/L) is a shorthand for Cyrilic and Latin respectively; and Scottish (G) abbreviates Gaelic.

signal provided by the multilingual NLI model.

4.1 Data Filtering

An intuitive approach to improving the factuality of machine generated summaries is to enhance the quality of the training data, simply by filtering noisy training samples (Nan et al., 2021a; Goyal and Durrett, 2021). More formally, given a training corpus $\mathbb{D}$ of input document-summary pairs, we find $\mathbb{D}^+ \subset \mathbb{D}$ such that for each document-summary pair $(d, s) \in \mathbb{D}^+$, $p(s \text{ is entailed by } d) > 0.5$.

We used our multilingual NLI model (see Section 3) to annotate the training data in XLSum (Hasan et al., 2021) for all 45 languages. Table 1 shows the total number training examples and the proportion where the summary was predicted to be entailed by the input (using a threshold of 0.5 on the NLI model score). The proportion of entailed summaries ranges from 68.96% (for Japanese) to 28.29% (for Punjabi). For all but three languages (English, Japanese, Nepali), the NLI model predicted *less than half* of the training summaries as being entailed by the input. We find these numbers strikingly low; this may be due to the nature of the dataset, since the relationship between news headlines and their corresponding article can be somewhat loose (e.g., headlines may include "clickbate" and additional details that are not mentioned in the article). Another reason might be errors of the NLI model; while it was shown to work well on TRUE/XNLI, XLSum may represent a different distribution.

Overall, the results in Table 1 indicate that filtering the training data based on the NLI signal can have a large impact on the resulting summarization model. Training on the entailed portion of the data may result in more factual summaries, however, at the expense of summary quality as the model unavoidably sees fewer examples (e.g., there are only 557 instances for Scottish Gaelic after filtering, while the original training set has 1,313).

4.2 Controlled Generation

Another way to leverage the NLI signal for improving the summarization model is via controlled generation (Keskar et al., 2019; Rashkin et al., 2021b). In this approach, special tokens are prepended to the model's input to indicate/control whether the output should be entailed or not.

Let $\mathbb{D}$ denote a training corpus of document summary pairs (d, s). We annotate each $(d, s) \in \mathbb{D}$ as (d', s), where d' is d prepeneded with an "<entailed>" symbol if $p(s \text{ is entailed by } d) > 0.5$, and otherwise d' is d prepended with "<not-entailed>". The model trained on $\mathbb{D}$ enhanced with these annotations is expected to learn the correlation between entailment and the special token value, and as a result to be "controlled" to produce more faithful summaries by prepending the token that corresponds to faithful (aka entailed) summaries in inference time. This method *implicitly* teaches the model to learn from the entailment signal while taking advantage of all available training data. It may, however, be more sensitive to wrong predictions by the entailment model as noisy examples are not discarded.

5 Experimental Setup

We focus on XLSum (Hasan et al., 2021), a recently introduced summarization dataset. XLSum extends to multiple languages the methodology put forward in Narayan et al. (2018a) for the creation of the English-only XSum; it contains 1 million BBC article-summary pairs covering 45 languages.

Model	Best-NLI	Best-ROUGE
Vanilla	3,600	15,000
Filtered	2,200	12,000
Controlled	3,400	8,800

Table 2: Number of finetuning steps for best checkpoint for each model according to NLI and ROUGE on the XLSum development set.

5.1 Model Details

We finetuned three models based on mT5 XXL (Xue et al., 2021, 14B parameters). The first is a "Vanilla" model which is trained on the XLSum data as-is. As previous work has shown that multilingual training improves performance for low-resource languages (Aharoni et al., 2019; Hasan et al., 2021), we also follow this setting and finetune a single massively multilingual model for all 45 languages in XLSum. The second model ("Filtered") is finetuned only on the portion of the data that passed the multilingual NLI filter. The third model ("Controlled") is trained on all data, using the controlled generation approach mentioned above. Specifically, for control tokens "<entailed>" and "<not-entailed>", we used two extra spare tokens from the mT5 vocabulary and prepended them to the input (Keskar et al., 2019; Rashkin et al., 2021b). During inference, we always prepend the input with "<entailed>" and report on the whole development and test sets.

Ideally, we would like to evaluate a *single* model checkpoint for *all* languages; in the literature, the best checkpoint is often selected using ROUGE. However, we also employ NLI scores to quantify improvements in faithfulness. For each model, we select two checkpoints that are best according to ROUGE and NLI (on the development set), when averaged across all languages. Table 2 summarizes the number of finetuning steps that led to the best checkpoints for each model according to ROUGE and NLI. For all models, we observe that best NLI checkpoints are earlier than ROUGE-based ones.

5.2 System Comparisons

We compare the above approaches to three additional baselines. Firstly, we record the number of examples that pass the NLI filter, per language, and select the same number at "Random". We then finetune a model similarly to the "Filtered" model above using this randomly selected data. Secondly, we introduce a "Self-ROUGE" baseline which selects examples where the ROUGE of the summary with respect to the input document is highest. Again, we choose the same number of examples as those which passed the NLI filter, and finetune a model on this data. Finally, we compare against model output from Hasan et al. (2021) who finetuned an mT5-Base pretrained model.

Quality: Is the summary comprehensible?

Incomprehensible: The summary is difficult to understand. It has serious grammatical errors, low fluency, and/or repeated information.
Somewhat Comprehensible: The summary makes sense but suffers from grammatical errors, low fluency, and/or repeated information.
Comprehensible: The summary is understandable. It does not exhibit any grammatical errors, disfluencies, and/or repeated information.

Attribution: Is all the information in the summary fully attributable to the article?

Yes, it is attributable: Select this option if it is accurate to say, "The provided news article says…" or "According to the news article…" with the summary following this phrase.
No, not fully attributable: Select this option if only some of the information is supported in the news article, but other parts of the information are missing from the news article or not an accurate representation.

Informativeness: Is the summary a good summary of the article?

Bad summary: The summary does not capture the important information in the article, or the captured information is not accurate with the article. It can also exhibit grammatical issues, low fluency, and/or repeated information.
Good Summary: The summary captures the important information in the article and presents it accurately and concisely. It does not exhibit any grammatical errors, disfluencies, and/or repeated information.

Figure 2: Instructions used in our human evaluation.

5.3 Automatic Evaluation

We report ROUGE (Lin, 2004) which is commonly used to measure the informativeness and fluency of model summaries against gold-standard references.[7] We also quantify faithfulness with the reference-free NLI score (Maynez et al., 2020; Honovich et al., 2022, *inter alia*). Since there are no tokenizers available for many of the languages in XLSum, we report ROUGE-L computed using the sentencepiece tokenization of mT5. Regarding NLI, we compute for each summary whether it is entailed by the input, and report the average over all examples in a partition (test or development set).

5.4 Human Evaluation

In addition to automatic evaluation, we conducted a large-scale human elicitation study assessing different dimensions of the output in all 45 languages. Firstly, we asked participants to read a system summary and assess its *Quality* (*Is the summary comprehensible?*), without looking at the source article, using a 1–3 rating scale where 3 means fully

[7]For ROUGE, we used the python implementation from https://github.com/google-research/google-research/tree/master/rouge

	ROUGE	NLI
Vanilla	33.65	64.40
Filtered $_{\text{best ROUGE}}$	**34.00**	72.17
Filtered $_{\text{best NLI}}$	32.98	**76.49**
Controlled $_{\text{best ROUGE}}$	33.00	72.17
Controlled $_{\text{best NLI}}$	33.28	71.38
mT5-Base (Hasan et al., 2021)	31.85	53.41
Self-ROUGE	33.12	67.39
Random	33.44	69.62

Table 3: ROUGE-L and NLI scores averaged across the 45 languages on the XLSum test set. Highest scores are in bold. Per language results are in Appendix C.

understandable and 1 indicates that the summary has serious fluency errors. After the first assessment, participants were shown the source article and were asked to rate the summary according to *Attribution* (*Is all the information in the summary fully attributable to the article?*) using the attribution definition[8] provided in Rashkin et al. (2021a); and *Informativeness* (*Is the summary a good summary of the article?*). Both assessments used binary judgements (we report on the percentage of times each system was rated positively). Figure 2 presents our instructions.[9]

In order to evaluate summarization output in such a diverse multilingual setting, we have taken several measures to scale our study to 45 languages while maintaining high inter-annotator agreement. We used the same instructions in English for all languages and invited bilingual participants (native speakers of the target language who are also proficient in English) to take part in our study.[10] Each participant had to pass a *Screener test* consisting of 25 questions with an accuracy of 85% before they could take part in the study. Finally, we conducted two pilot studies before the final evaluation to give participants feedback and improve agreement. Our final elicitation study was conducted using 100 instances per language, each randomly sampled from the test set. We collected ratings from three different annotators for each data point.

6 Results

Filtered Model is Best on NLI-based Evaluation Table 3 presents our results on the test set averaged across all 45 languages, for our three model variants (Vanilla, Filtered, and Controlled) and three baselines. For the Filtered and Controlled models we report results for both the best-ROUGE and best-NLI checkpoints, while for the others we only use ROUGE for checkpoint selection as no NLI model is involved in their training. Per language results on the validation and test sets are in Appendix C, Tables 9 and 10.[11]

We see that the Filtered model outperforms all other models across languages achieving an average score of 76.49 for the Best-NLI checkpoint (it obtains best NLI scores in 43 out of 45 languages). This suggests that data filtering is a viable approach for improving the factual consistency of summarization systems. The next best models in terms of NLI are the Filtered and Controlled variants (with Best-ROUGE checkpoints), achieving an average score of 72.17. The Controlled and Vanilla models perform mostly worse than the Filtered variant in terms of NLI with either Best-NLI or Best-ROUGE checkpoints. Note the significant NLI score gap between the Vanilla model and the Best-NLI Filtered model (12.19 points on average). This primarily points to the quality of the unfiltered data, since both models are based on T5-XXL. The Best-NLI checkpoint outperforms the Best-ROUGE checkpoint for the Filtered model (average NLI scores of 76.49 vs 72.17). However, we observe a degradation of 0.79 NLI points when comparing the Best-NLI and Best-ROUGE checkpoints for the Controlled model.

Effect of Entailment Signal on ROUGE As shown above, NLI scores improve in all languages when training uses the signal from the NLI model, either by filtering data or by using controlled generation. But what is the effect on ROUGE? Looking at the average ROUGE scores across languages in Table 3, we again see that the best ROUGE is obtained by the Filtered model, with the Best-ROUGE checkpoint. Interestingly, this model is trained on much fewer examples, but obtains better results than the Vanilla and Controlled variants that use all training examples in XLSum. This model obtains higher or comparable NLI scores (72.17 and 76.49, for Best-ROUGE and Best-NLI, respectively) than the other models, suggesting that it is more accurate with respect to the reference summaries and more faithful with respect to the input. In general, the Vanilla, Filtered and Controlled models obtain very

[8] A fully attributable (or supported) system-generated summary contains an accurate representation of information in the source news article. No information in the summary is unattested when compared against the source news article.

[9] We also present an example of the interface presented to our participants in Appendix D (Figure 3).

[10] See Appendix D for more information on annotators' qualifications and demographic information.

[11] Due to the high computational cost of the experiments, our results are based on a single run.

	Vanilla				Filtered				Controlled			
	Best-ROUGE		Best-NLI		Best-ROUGE		Best-NLI		Best-ROUGE		Best-NLI	
Language	ROUGE	NLI	ROUGE	NLI	ROUGE	NLI	ROUGE	NLI	ROUGE	NLI	ROUGE	NLI
English	32.51	68.31	32.93	74.23	**33.23**	80.32	32.4	**84.4**	33.07	81.75	33.14	82.99
Training Resources												
High	33.21	63.42	**33.25**	67.26	32.29	**69.24**	31.23	67.79	28.53	64.79	29.37	64.80
Medium	33.21	63.22	32.37	66.37	**33.46**	69.68	32.54	**74.09**	32.64	69.96	32.39	67.52
Low	**34.38**	65.89	33.69	69.00	34.03	71.98	33.03	**77.61**	33.65	71.88	33.69	71.40
Language Family												
Indo-European	32.62	63.23	32.11	66.22	**33.07**	69.44	32.32	**74.48**	32.02	69.61	32.20	68.18
Romance	31.61	57.03	31.04	61.10	**31.91**	66.51	31.27	**70.21**	31.00	68.83	31.48	66.09
Turkic	28.90	64.87	29.32	65.50	**29.38**	70.32	28.85	**76.35**	28.25	71.19	28.99	70.18
Semitic	34.78	65.85	34.65	72.02	**34.95**	74.73	33.87	**77.62**	34.13	73.00	34.60	74.36
Afro-Asiatic	**33.49**	62.58	32.10	67.31	32.75	71.08	32.03	**76.07**	32.71	68.31	32.53	68.89
Indo-Iranian	36.77	67.55	37.07	68.14	**38.04**	73.39	37.10	**76.19**	35.98	70.77	36.77	71.84
XNLI Training Data												
Available	33.01	62.68	33.09	67.30	**34.03**	73.53	33.06	**77.15**	32.40	73.78	33.48	72.71
Unavailable	33.90	64.97	33.21	68.03	**33.99**	71.63	32.94	**76.23**	33.27	71.52	33.20	70.83

Table 4: ROUGE-L and NLI scores on XLSum test set for best checkpoints averaged across language groups. For training resources we consider three groups with varying numbers of training examples: High ([70K–10K]), Medium ([10K–6K]), and Low (less than 6K). For language families, the Indo-European cluster represents Bengali, Gujarati, Hindi, russian, Serbian (Cyrillic and Latin), and Sinhala; the Romance cluster comprises of French, Protuguese, and Spanish; the Turkic cluster contains Azerbaijani, Kyrgyz, Turkish, and Uzpek; Semitic languages are Amharic, Arabic, and Tigrinya; the Afro-Asiatic cluster groups together Hausa, Oromo, and Somali; finally, the Indo-Iranian cluster represents Pashto, Persian, and Punjabi; we omit clusters with two members and singletons. We also create two subsets depending on whether they appear in the XNLI dataset used to train our multilingual NLI model (Available, Section 3) or not (Unavailable). Highest ROUGE-L and NLI numbers are in bold.

similar ROUGE scores, ranging between 32.98 and 34.00, while the range of NLI scores is much larger (from 64.30 to 76.49).

Comparison against Baseline Approaches Table 3 also compares to previous work (mT5-Base, Hasan et al. 2021), and the Self-ROUGE and Random selection baselines. We did not employ any NLI preprocessing in building the baseline models, neither in filtering or checkpoint selection. We observe that all model variants (Vanilla, Filtered, and Controlled) are superior to mT5-Base in terms of ROUGE which is not surprising given the different model capacities (XXL vs Base). We also see that *any* filtering improves NLI scores (compare Vanilla against Self-ROUGE and Random), incurring a slight decrease in terms of ROUGE, while *targeted* filtering using NLI yields best results.

ROUGE and NLI across Different Language Groups Table 4 shows our results clustered by (1) the number of training examples per language: High (10k–70k examples), Medium (6k–10k examples); and Low (less than 6k examples); (2) language family (Indo-European, Romance, Turkic, Semitic, Afro-Asiatic, and Indo-Iranian families); and (3) whether XNLI training data is available; we cluster languages into two subsets, those that appear in the XNLI dataset used to train our NLI model, and those that do not (see Table 1). We report results on English on its own, as it is the language with the largest number of examples (370k).

Again, we observe that the Filtered model is in most cases superior, including English. Vanilla scores are better on ROUGE for Low resource and Afro-Asiatic languages, although the difference against other models is less than 1 ROUGE point. The Controlled model is not better than Filtered or Vanilla in any configuration, irrespective of how languages are grouped into clusters. In conclusion, we find that the Filtered model dramatically improves faithfulness, while maintaining ROUGE performance similar to other models. We present examples of model output in Appendix G.

Human Assessment for Quality, Attribution, and Informativeness Table 5 presents our human evaluation results for Quality, Attribution and Informativeness (it also includes automatic evaluation results for a side-by-side comparison). We provide per language analysis in Appendix E (See Tables 15–17) and aggregate statistics using the same groups as in Table 4 (see Tables 18–20).

Unsurprisingly, human reference summaries were more understandable than Vanilla, Filtered, or Controlled summaries, with least fluency issues. Differences between the gold standard summaries and those generated by the Filtered Best-NLI and Controlled Best-NLI are, however, not statistically significant (using a one-way ANOVA with post-hoc Tukey HSD tests; $p < 0.01$). Summaries generated using our Filtered Best-NLI model were most attributed (or faithful) and informative, with respect to their input documents. Differences be-

tween the Filtered Best-NLI model and all other comparisons are statistically significant (using a one-way ANOVA with post-hoc Tukey HSD tests; $p < 0.01$). In conclusion, human evaluation confirms the Filtered model is best at generating faithful and informative summaries.

Effect on Summary Length One may argue that we are improving faithfulness by favoring shorter summaries. To study this, we also report in Table 5 the ratio of predicted to target summary length averaged across all test examples, for different models.

As we can see, best-NLI checkpoints do yield a reduction in predicted length across different models compared to their Best-ROUGE checkpoints; the length ratios drop from 0.93 to 0.88 for Vanilla, from 0.89 to 0.87 for Filtered, and from 1.00 to 0.81 for Controlled. However, shorter summaries are not necessarily more faithful; the worst length ratio (0.81) is for the Controlled Best-NLI model which performs worse on NLI, Attribution, and Informativeness, compared to the Filtered Best-NLI model with a higher length ratio (0.87). The Filtered Best-NLI model only yields a marginal reduction in summary length compared to the Vanilla Best-Rouge summaries (Length ratio: 0.87 vs 0.93), but improves on NLI scores (76.50 vs 64.31), Quality (0.86 vs 0.85), Attribution (0.52 vs 0.44), and Informativeness (0.45 vs 0.37) assessments.

7 Metric Learning for Multilingual Summary Evaluation

Our large-scale judgment elicitation study (across multiple languages and system outputs) delivered valuable annotations of summary document-quality (31,499 pairs x 3 quality dimensions x 3 raters). We next explore whether it is possible to *learn* metrics for evaluating Quality, Attribution, and Informativeness automatically. Existing metrics (e.g., BLEURT; Sellam et al. 2020) have not targeted summarization specifically, or considered attribution, and multiple languages. Let $\boldsymbol{s} = (s_1, \ldots, s_r)$ denote a summary of length r where each s_i is a token and let $\boldsymbol{d} = (d_1, \ldots, d_p)$ be its corresponding input document of length p. Let $\{(d_i, x_i, y_i)\}_{n=1}^{N}$ be a training dataset of size N where $y_i \in R$ is the human rating that indicates how good x_i is as a summary of d_i along a specific dimension. Our goal is to learn a function $\boldsymbol{f} : (\boldsymbol{d}, \boldsymbol{x}) \rightarrow y$ that predicts the human rating.

We finetuned three models based on mT5-XXL (Xue et al., 2021), one per dimension (details in Appendix F). The input was the concatenation of a document and its summary, and the output the human rating. 10% of the elicited ratings (across languages) were reserved for testing, while the remainder was used for training and validation. Table 6 reports correlation coefficients (Pearson's r) between model predictions and (mean) human ratings. MT5-Q, MT5-A and MT5-I denote the learned metrics corresponding to Q(uality), A(ttribution), and I(informativeness), respectively. In addition, we report correlation coefficients for ROUGE and NLI.

Overall, we observe that learned metrics correlate best with human ratings (across dimensions). ROUGE correlates weakly with human judgments but cannot distinguish any dimension in particular, whereas NLI scores reliably correlate with attribution. Our results underscore the need for better and more fine-grained evaluation of summary quality, and also corroborate well-known issues (Gehrmann et al., 2022) with widely adopted lexical overlap-based metrics such as ROUGE.

8 Conclusion

In this paper we leveraged factual consistency evaluation for improving summarization models in multiple languages. Extensive experiments on the XL-Sum dataset showed large gains when training summarization models on a subset of the data selected using the NLI signal. Through a large-scale human evaluation study, we obtained ratings which not only helped us distinguish best performing systems, but were further used to learn metrics for assessing multilingual summaries along the dimensions of Quality, Attribution, and Informativeness. These metrics could be further used to inspect the quality of summarization datasets. Our annotators found that summaries are (on average) only 52% of the time fully faithful to their documents and this number is much worse for some languages (e.g., Hausa, Yoruba; see Table 16 in Appendix E).

An interesting avenue for future is to directly optimize the summarization models towards the different quality objectives, e.g. via Reinforcement Learning (Narayan et al., 2018b) or Calibrating Sequence Likelihood (Zhao et al., 2022).

Limitations

While our work covers a large number of languages, it is focused on a specific source and style of summaries. Our experiments focus exclusively on the XLSum dataset (Hasan et al., 2021) which is based

Metric	Vanilla		Filtered		Controlled		Reference
	Best-ROUGE	Best-NLI	Best-ROUGE	Best-NLI	Best-ROUGE	Best-NLI	
Quality	0.85	0.84	0.85	0.86	0.84	0.86	**0.88**
Attribution	0.44	0.47	0.46	**0.52**	0.49	0.47	0.31
Informativeness	0.37	0.40	0.39	**0.45**	0.41	0.40	0.27
Length Ratio	0.93	0.88	0.89	0.87	**1.00**	0.81	1.00
ROUGE	33.65	33.18	**34.00**	32.98	33.02	33.28	—
NLI	64.31	67.82	72.18	**76.50**	72.18	71.38	—

Table 5: Mean human judgments on XLSum test set averaged across languages. We also include ROUGE-L and NLI scores for a side-by-side comparison. Length Ratio is the ratio of predicted length to target length averaged across all test examples. Best results in each row are in **bold**.

Metric	Quality	Attribution	Informativeness
ROUGE	0.12	0.15	0.12
NLI	0.08	0.36	0.33
MT5-Q	**0.38**	0.09	0.14
MT5-A	0.09	**0.57**	0.52
MT5-I	0.14	0.49	**0.53**

Table 6: Correlation of metrics with human summary ratings for the dimensions of Quality (Q), Attribution (A), and Informativeness (I) on the test set. All correlations are statistically significant at $p < 0.01$.

on BBC articles where the opening sentence serves as a summary. It would be interesting to explore our methods on additional datasets and text generation tasks, e.g., where the summaries are longer, or there are multiple input documents.

Ethics Statement

An ethical consideration that concerns our work is the problem of misinformation. While we make a step towards improving the factual consistency of text generation systems which in turn should alleviate issues of misinformation, it is important to note that current systems are still far from being perfect in this respect, and thus should be used with caution.

Acknowledgements

We thank Ankur Parikh, Sebastian Gehrmann, Dipanjan Das and William Cohen for their feedback on this work. The human rating process was managed by Muqthar Mohammad, Kiranmai Chennuru, Aishwarya Gomatam, Raghava Ram Pamidigantam and Mahesh Maddinala, without them this work would not have been possible. Thanks for invaluable support from Sheila de Guia and Suneet Dhingra.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. 2021. Focus attention: Promoting faithfulness and diversity in summarization. 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 6078–6095, Online. Association for Computational Linguistics.

Kristjan Arumae and Fei Liu. 2019. Guiding extractive summarization with question-answering rewards. 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 2566–2577, Minneapolis, Minnesota. Association for Computational Linguistics.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel 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, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

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.

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.

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

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *CoRR*, abs/2112.08542.

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.

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, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati 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 Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, 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.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text.

Navita Goyal, Balaji Vasan Srinivasan, Anandhavelu N, and Abhilasha Sancheti. 2021. Multi-style transfer with discriminative feedback on disjoint corpus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3500–3510, Online. Association for Computational Linguistics.

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

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Ranit Aharonov, and Sachindra Joshi. 2021. Using question answering rewards to improve abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 518–526, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi 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*, pages 4693–4703, Online. Association for Computational Linguistics.

Or Honovich, Roee 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 Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee 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*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *ArXiv*, abs/1909.05858.

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*.

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

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

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.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021a. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021b. Improving factual consistency of abstractive summarization via question answering. 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 6881–6894, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. 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.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins, and Mirella Lapata. 2022. A well-composed text is half done! composition sampling for diverse conditional generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 1319–1339, Dublin, Ireland. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

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.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 646–653, New Orleans, Louisiana. 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. *CoRR*, abs/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.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, et al. 2022. Scaling up models and data with t5x and seqio. *arXiv preprint arXiv:2203.17189*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *CoRR*, abs/2002.08910.

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching sentence-pair nli models to reason over long documents and clusters. *arXiv preprint arXiv:2204.07447*.

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. 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.

Kaiqiang Song, Lin Zhao, and Fei Liu. 2018. Structure-infused copy mechanisms for abstractive summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1717–1729, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryściński, Justin F. Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors.

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. 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.

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*, pages 483–498, Online. Association for Computational Linguistics.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.

Yao Zhao, Misha Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J. Liu. 2022. Calibrating sequence likelihood improves conditional language generation. *ArXiv*.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.

A Intrinsic NLI Model Evaluation

In this section we present more detailed evaluation results for our our multilingual NLI model. Table 7 shows accuracy the on XNLI test set for 15 languages; we compare our results to the model of Xue et al. (2021); it is also based on mT5-XXL finetuned on MNLI. The table shows our model achieves higher accuracy for all languages, however, results are not fully comparable as we only consider binary labels (entailment/non-entailment) in our setting in comparison to three classes (entailment/neutral/contradiction) for Xue et al. (2021).

We additionally evaluate our model on the TRUE factual consistency benchmark (Honovich et al., 2022). TRUE consists of 11 diverse datasets (including the output of grounded text generation systems), annotated with binary factual consistency labels. Although TRUE only includes English examples, we use it for our evaluation due to its relevance to factual consistency in summarization. Table 8 shows the area under the ROC curve (ROC AUC) results for all dataset in TRUE where we compare our multilingual model to T5-11B trained on ANLI (Nie et al., 2020) reported by Honovich et al. (2022). Results show that our multilingual model performs on par with theirs, while finetuning our model on non-English data causes a slight decrease in performance.

B Technical Modeling Details

We used the *t5x* (Roberts et al., 2022) framework for all training and inference tasks. We ran all experiments on TPU accelerators.

C Detailed Automatic Evaluation Results

Table 9 (development set) and Table 10 (test set) report ROUGE-L and NLI scores on XLSum broken for individual languages. Table 11 compares our Filtered model against previous work (mT5-Base, Hasan et al. 2021) and the Self-ROUGE and Random selection baselines. Results are presented for individual languages and on average on the XLSum *test set*.

Table 12 shows details of how we grouped languages into different clusters (i.e., family and the availability of NLI training data). Table 13 shows our results on the XLSum *development set* clustered by (1) the number of training examples per language; we group languages into three clusters – High (10k–70k examples), Medium (9k–16k examples); and Low (less than 6k examples); (2) language family (we group languages into Indo-European, Romance, Turkic, Semitic, Afro-Asiatic, and Indo-Iranian families); and (3) whether XNLI training data is available; we cluster languages into two subsets, those that appear in the XNLI dataset used to train our multilingual NLI model, and those that do not (see Table 1). We report results on English on its own, as it is the language with the largest number of examples (370k).

D Human Evaluation Setup and Annotator Qualifications

Figure 3 presents a snapshot of the interface seen by our participants together with the instructions used in our human evaluation studies.

To recruit our participants, we screened their language skills to determine whether they're native speakers, their education level and country of residence as well as origin. For some languages we could not recruit native speakers in the country of birth for various restrictions and sourcing difficulties, we hired native speakers in other countries. In addition, we created a screener test to determine the raters' suitability for the task. In total, we recruited 388 raters across all 45 locales. 2.58% of them hold a Doctorate, 31.96% holds a master degree, 57.73% of them hold a bachelor degree, 7.73% hold High school degree or equivalent. Table 14 presents the demographics of our participants. All our annotators are paid adequately by our suppliers adhering to the supplier code of conduct.

E Detailed Human Evaluation Results

Table 18 presents human evaluation results for Summary Quality for individual languages on the XL-Sum test set. Table 16 shows mean judgments for Attribution, again per language, and finally Table 16 summarizes our results for Informativeness. We also group human judgments according to number of training examples, language family, and whether XNLI training data is available. Tables 18–20 show these different types of clustering for the judgments pertaining to Summary Quality, Attribution, and Informativeness.

F Metric Training Details

The metrics were trained by finetuning mT5-XXL to predict a binarized version of the human judgments (a summary receives a score of 1 if the mean human rating > 0.5). Each metric is trained for 20,000 steps with batch size $= 32$ and a learning

Model	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg.
(Xue et al., 2021)	84.5	87.7	87.3	87.3	91.6	87.8	86.9	83.2	85.1	80.3	81.7	83.8	79.8	84.6	83.6	85.0
Ours	90.0	91.6	91.2	91.4	93.5	91.7	91.4	88.7	90.3	87.2	88.0	89.1	86.3	89.6	89.8	90.0

Table 7: Accuracy results on XNLI test set.

Model	FRANK	MNBM	QAGS-C	QAGS-X	BEGIN	Q^2	DialFact	PAWS	FEVER	VitaminC	SummEval	Avg.
(Honovich et al., 2022)	89.4	77.9	82.1	83.8	82.6	72.7	77.7	86.4	93.2	88.3	80.5	83.1
Ours	88.1	77.2	82.8	84.5	80.5	72.8	80.1	82.6	93.7	88.2	75.6	82.4

Table 8: ROC AUC results on TRUE.

rate of 0.0001. Checkpoints were selected by their accuracy on a validation set.

G Example Output

We showcase summaries generated by our models in Tables 21, 22, and 23. The article in Table 21 discusses a cholera outbreak in Algeria, with two deaths, 46 confirmed cases and 88 suspected cases. The reference summary in addition mentions that there have been 139 hospitalizations since August 2018, however, the number of hospitalizations is not given in the input document. Tha Vanilla summary manages to hallucinate two facts: the deaths have not been *several*, they are only two, and the number of suspected cases is 88, not 100. The Controlled summary is factual although perhaps sparse with the details, it only mentions the cholera deaths but not the cases. The Filtered summary on the other hand correctly mentions the number of confirmed and suspected cases but does not mention the deaths.

The article in Table 22 talks about trials of an Ebola vaccine in Oxford. The trial involves 72 volunteers, and preliminary tests on monkeys have shown that the vaccine confers immunity against Ebola. Similar small-scale trials are underway in the United states and three African countries spared from the epidemic. The Reference summary is factually correct, the Vanilla summary gives the false impression that the Ebola trial is large-scale; by omitting the adjective "large-scale", the Controlled summary is factual, and likewise the Filtered summary does not include any hallucinations.

The article in Table 23 talks about Greenpeace activists arrested in Russia on piracy charges for protesting against and oil rig in the arctic sea. Among the 30 arrested, two are Argentinian, one Brazilian. Five of them were accused of climbing the oil rig and their detention was extended by two months. The Reference summary is factually correct, the Vanilla summary has slight fluency issues, it hallucinates the location of the oil rig to be in the Black Sea, and misrepresents the protest to be about the oil rig closure; the Controlled summary is factual but focuses on the oil rig climbers, and likewise the Filtered summary does not contain any hallucinations but focuses on the Argentine activists only.

	Vanilla				Filtered				Controlled			
	Best-ROUGE		Best-NLI		Best-ROUGE		Best-NLI		Best-ROUGE		Best-NLI	
Language	**ROUGE**	**NLI**	**ROUGE**	**NLI**	**ROUGE**	**NLI**	**ROUGE**	**NLI**	**ROUGE**	**NLI**	**ROUGE**	**NLI**
Amharic	33.54	60.4	33.02	65.43	33.17	66.47	32.72	**73.38**	**33.57**	62.12	32.94	66.09
Arabic	29.94	38.2	31.94	53.01	**32.49**	61.29	32.11	**65.29**	30.58	52.63	31.91	58.97
Azerbaijani	**29.05**	53.14	28.76	51.56	28.79	61.14	28.39	**67.39**	27.84	53.96	28.25	55.29
Bengali	**33.96**	62.85	33.22	61.61	33.36	70.6	33.61	**77.53**	33.68	59.77	32.88	67.1
Burmese	**40.88**	57.62	39.6	60.48	39.19	70.19	38.91	**76.96**	39.41	62.03	39.4	66.49
Chinese (simp.)	32.44	47.32	32.76	58.2	**34.93**	71.09	34.31	**72.99**	32.61	62.88	33.64	67.46
Chinese (trad.)	33.57	52.99	34.11	58.7	**36.3**	**72.42**	34.98	71.09	33.28	62.3	34.35	66.06
English	31.68	55.52	32.26	70.7	**32.37**	72.83	32.3	**80.48**	32.61	69.61	32.44	74.58
French	**33.62**	51.19	33.04	51.02	33.13	61.24	32.94	**64.92**	32.6	60.07	32.82	58.26
Gujarati	**32.83**	50.1	31.93	48.79	32.04	55.54	32.25	**61.14**	32.45	53.92	31.59	52.64
Hausa	**37.34**	48.85	35.64	51.52	36.74	56.83	35.96	**62.81**	36.93	54.71	35.89	57.34
Hindi	35.53	56.8	35.67	62.27	**36.39**	**68.92**	36.1	68.31	35.62	62.78	36.15	62.02
Igbo	35.09	44.85	35.23	48.13	35.36	57.84	35.14	**59.15**	34.98	48.25	**35.6**	52.89
Indonesian	33.06	57.62	33.32	60.29	**34.28**	69.7	33.32	**70.54**	33.3	63.68	33.38	68.13
Japanese	40.3	68.48	39.82	72.57	**40.9**	76.52	39.7	77.08	39.92	**81.95**	39.54	81.07
Kirundi	32.37	47.9	31.32	46.54	32.44	52.65	31.2	**60.33**	**33.15**	49.56	31.74	51.04
Korean	39.83	66.88	37.21	67.36	38.96	70.23	37.75	**79.42**	**39.89**	70.31	36.29	74.05
Kyrgyz	25.8	53.27	**26.37**	50.89	24.78	56.52	25.29	**68.1**	25.29	58.87	25.62	60.43
Marathi	28.62	47.15	27.2	52.72	30.09	**63.59**	**30.84**	63.35	29.32	49.29	26.81	52.68
Nepali	37.96	67.53	37.48	67.11	**38.14**	72.3	37.06	**74.48**	37.4	67.72	36.32	71.51
Oromo	**29.63**	59.69	27.37	63.12	28.83	68.41	28.07	**75.19**	28.61	59.08	28.19	61.31
Pashto	38.96	58.37	38.57	57.27	**39.55**	70.66	38.73	**73.06**	38.19	60.74	38.17	65.73
Persian	34.91	60.31	36.35	64.62	**37.37**	71.47	36.83	**72.66**	35.55	63.05	35.99	68.45
Pidgin	34.27	51.04	33.78	60.66	**34.5**	67.76	33.27	**67.96**	34.43	57.5	34.1	61.76
Portuguese	32.33	39.16	32.21	48.18	**33.43**	57.06	32.66	**57.37**	32.4	48.76	32.67	57.22
Punjabi	**34.88**	48.49	33.82	48.97	34.41	50.05	34.53	**55.12**	34.46	45.76	33.87	50.55
Russian	27.33	47.09	28.02	52.99	**28.06**	56.07	27.94	**61.11**	27.89	59.66	27.95	61.59
Scottish Gaelic	**33.7**	50.55	33.36	61.03	32.38	56.82	32.63	**72.31**	33.24	63.18	32.71	62.77
Serbian (Cyrillic)	**27.76**	47.21	27.14	48.95	27.69	55.19	26.78	**57.38**	27.34	50.49	26.55	51.76
Serbian (Latin)	27.15	38.59	26.84	47.07	**27.95**	50.49	26.25	**55.74**	27.14	44.18	25.19	41.25
Sinhala	**36.16**	57.9	34.36	60.04	34.67	61.27	34.24	**76.16**	35.3	66.45	34.61	69.81
Somali	30.93	51.75	30.46	60.48	**31.57**	62.42	30.52	**69.28**	30.9	57.59	30.76	57.92
Spanish	26.63	38.53	26.25	42.13	**26.88**	48.84	26.48	**61.87**	26.67	46.36	26.65	44.34
Swahili	**34.95**	56.8	34.83	60.1	34.91	66.87	33.66	**71.64**	34.74	62.74	33.61	61.51
Tamil	30.82	61.6	29.5	70.35	**31.98**	78.94	30.46	**81.49**	30.22	75.1	29.81	75.82
Telugu	27.9	54.06	27.94	56.94	28.01	61.29	28.29	**69.97**	**28.47**	54.84	27.53	58.79
Thai	29.46	54.3	**30.22**	61.89	29.2	68.32	29.19	**72.93**	29.08	60.84	29.78	65.0
Tigrinya	**35.25**	57.08	33.9	63.66	35.11	67.61	32.9	**73.36**	33.41	52.28	34.69	60.57
Turkish	29.94	50.76	30.75	54.47	**32.18**	66.95	31.39	**68.55**	30.64	57.64	31.09	60.72
Ukrainian	28.16	44.18	28.6	52.46	**28.75**	61.42	27.98	**63.55**	28.43	56.66	28.49	60.08
Urdu	36.03	52.54	35.79	52.69	**37.34**	66.87	36.75	**70.13**	36.02	59.9	36.15	62.75
Uzbek	28.1	60.46	27.94	59.44	**28.12**	61.3	27.75	**70.78**	27.7	59.99	27.46	59.42
Vietnames	34.71	53.8	34.65	59.88	**35.48**	67.14	34.95	**68.21**	34.38	59.73	35.1	64.89
Welsh	**34.56**	55.44	32.7	52.02	33.64	**66.7**	32.21	62.06	33.5	57.04	32.46	60.19
Yoruba	**37.25**	54.79	35.76	54.26	36.66	58.31	36.04	**66.82**	36.82	50.83	36.03	55.87
Average	32.87	53.18	32.47	57.16	**33.17**	63.91	32.56	**68.65**	32.67	58.59	32.38	61.43

Table 9: ROUGE-L and NLI scores per language on the XLSum development set for the Best-ROUGE and Best-NLI checkpoints (chosen by averaging across all languages). Highest scores in each row are in **bold**.

	Vanilla				Filtered				Controlled			
	Best-ROUGE		Best-NLI		Best-ROUGE		Best-NLI		Best-ROUGE		Best-NLI	
Language	**ROUGE**	**NLI**	**ROUGE**	**NLI**	**ROUGE**	**NLI**	**ROUGE**	**NLI**	**ROUGE**	**NLI**	**ROUGE**	**NLI**
Amharic	**35.60**	72.56	34.91	78.71	35.15	74.65	34.55	**79.74**	35.1	75.52	35.22	79.12
Arabic	32.00	55.15	33.31	64.02	33.33	72.83	33.05	**74.88**	31.67	71.11	**33.34**	71.68
Azerbaijani	**29.95**	62.15	29.07	62.2	29.28	68.03	29.26	**73.41**	27.99	66.75	29.0	67.03
Bengali	**34.19**	71.12	33.71	74.49	34.17	76.53	33.38	**81.14**	34.14	73.64	33.81	73.48
Burmese	**41.40**	68.54	39.83	69.56	40.58	73.68	39.08	**76.26**	40.55	70.74	40.78	74.32
Chinese (simp.)	33.54	62.43	33.08	69.07	**35.66**	76.95	34.15	**81.3**	32.48	78.83	34.44	77.72
Chinese (trad.)	34.39	63.10	34.37	69.02	**36.49**	78.15	34.43	**79.48**	32.86	77.50	34.95	77.95
English	32.51	68.31	32.93	74.23	**33.23**	80.32	32.4	**84.4**	33.07	81.75	33.14	82.99
French	34.12	64.88	33.87	67.61	**34.49**	72.94	33.98	**76.4**	33.18	77.11	34.35	72.59
Gujarati	**33.21**	62.52	32.19	62.12	32.59	63.54	32.46	**71.64**	32.23	62.38	32.74	62.85
Hausa	**38.02**	60.59	36.51	63.07	37.3	67.47	36.75	**72.18**	37.25	66.37	36.97	64.58
Hindi	36.88	66.74	36.92	69.77	**37.45**	74.20	36.97	**77.26**	36.17	73.31	37.26	74.70
Igbo	34.17	56.01	**35.45**	60.22	35.44	64.36	34.19	**72.00**	35.3	66.77	35.32	70.01
Indonesian	34.47	67.63	34.05	72.5	**35.26**	79.06	33.76	**82.31**	34.05	78.27	34.66	78.65
Japanese	41.19	76.81	40.86	79.35	**41.30**	78.60	39.52	81.29	41.0	**87.84**	40.65	85.60
Kirundi	**33.70**	65.50	32.67	66.31	33.63	70.66	31.74	**75.37**	33.05	72.57	32.73	67.72
Korean	**40.36**	72.05	39.28	74.00	39.83	78.51	38.24	**82.18**	39.9	80.02	39.28	78.12
Kyrgyz	26.48	67.57	**27.02**	65.58	26.54	67.99	25.41	**78.02**	26.82	73.01	26.70	73.39
Marathi	30.11	61.84	28.08	64.23	**31.31**	67.86	30.95	**70.41**	28.85	66.95	27.75	63.77
Nepali	**39.25**	76.43	38.8	78.46	38.67	80.57	37.48	**85.65**	38.42	80.86	37.40	77.18
Oromo	**30.67**	68.92	28.76	72.55	29.61	76.35	28.76	**79.62**	29.53	71.34	29.62	73.94
Pashto	39.07	70.01	39.04	70.9	**39.88**	76.34	38.83	**80.21**	38.45	72.46	38.85	73.72
Persian	35.11	71.68	36.47	71.57	**37.86**	79.22	36.73	**80.7**	34.16	78.50	35.82	80.23
Pidgin	34.33	62.07	32.95	67.91	**34.41**	72.03	33.07	**76.32**	34.07	70.35	33.82	69.85
Portuguese	33.3	52.89	32.88	57.12	**33.9**	64.24	33.14	64.91	32.76	68.01	33.16	**68.12**
Punjabi	36.12	60.96	35.71	61.95	**36.39**	64.61	35.73	**67.67**	35.33	61.34	35.64	61.57
Russian	28.48	58.42	28.38	62.88	**28.71**	69.20	28.26	**72.51**	28.09	71.88	28.58	70.54
Scottish Gaelic	33.41	57.38	33.01	66.48	32.95	68.14	32.71	**76.06**	**33.47**	71.90	32.57	66.07
Serbian (Cyrillic)	28.31	56.71	27.32	62.36	**29.20**	62.97	27.93	**69.32**	28.14	66.63	27.68	62.71
Serbian (Latin)	26.67	51.23	25.49	54.11	**28.21**	59.93	26.84	**65.08**	25.74	62.00	24.09	51.26
Sinhala	**36.33**	71.25	35.72	74.55	36.04	74.76	35.67	**81.58**	35.53	73.61	36.03	74.28
Somali	**31.77**	58.23	31.04	66.32	31.35	69.42	30.59	**76.42**	31.36	67.22	30.99	68.14
Spanish	27.40	53.33	26.38	58.58	**27.33**	62.34	26.7	**69.32**	27.06	61.36	26.92	57.55
Swahili	35.64	64.57	35.93	69.28	**36.46**	73.25	35.23	**78.58**	35.59	73.35	36.05	68.49
Tamil	31.94	70.37	30.31	77.9	**33.26**	81.91	30.99	**83.11**	31.49	82.14	30.93	81.76
Telugu	**29.35**	60.61	28.61	67.74	29.11	67.83	28.88	**75.10**	29.11	66.30	28.4	66.54
Thai	30.59	61.83	29.92	68.77	**30.66**	73.27	29.63	**78.95**	29.1	74.01	30.42	71.33
Tigrinya	**36.73**	69.83	35.73	73.32	36.37	76.71	34.01	**78.24**	35.63	72.37	35.24	72.27
Turkish	30.80	62.36	32.19	64.58	**33.65**	74.66	32.45	**75.65**	30.81	74.60	32.5	72.57
Ukrainian	28.34	58.83	28.61	63.15	**28.99**	70.87	27.66	**73.64**	28.01	72.41	28.42	72.61
Urdu	36.93	67.84	37.12	69.5	**38.18**	74.81	37.05	**77.31**	36.14	73.41	37.45	75.64
Uzbek	28.37	67.41	**29.02**	69.66	28.06	70.62	28.27	**78.32**	27.39	70.40	27.76	67.74
Vietnames	35.91	65.94	35.77	67.54	**36.69**	73.01	35.46	**76.94**	34.99	70.91	35.86	71.52
Welsh	**35.31**	65.27	33.75	65.32	33.89	76.06	32.47	**76.11**	34.33	71.58	33.66	71.47
Yoruba	**37.65**	63.91	36.01	63.22	37.01	68.48	35.15	**75.41**	35.63	68.52	36.78	68.55
Average	33.65	64.30	33.18	67.82	**34.00**	72.17	32.98	**76.49**	33.00	72.17	33.28	71.38

Table 10: ROUGE-L and NLI scores per language on the XLSum test set for Best-ROUGE and Best-NLI checkpoints. Highest scores in each row are in **bold**.

	Filtered		XLSum mT5-Base		Self-ROUGE		Random	
Language	**ROUGE**	**NLI**	**ROUGE**	**NLI**	**ROUGE**	**NLI**	**ROUGE**	**NLI**
Amharic	35.15	74.65	31.14	55.07	33.76	74.96	34.33	76.97
Arabic	33.33	72.83	32.64	53.90	34.35	67.29	34.19	69.55
Azerbaijani	29.28	68.03	27.14	48.94	28.04	61.06	26.83	66.14
Bengali	34.17	76.53	30.90	61.05	32.75	73.34	34.04	75.61
Burmese	40.58	73.68	36.64	51.81	38.63	60.56	37.87	73.03
Chinese (simp.)	35.66	76.95	37.86	61.61	36.22	70.42	36.36	76.37
Chinese (trad.)	36.49	78.15	37.78	62.45	36.70	71.09	36.66	77.58
English	33.23	80.32	34.16	62.83	37.21	75.68	37.46	77.25
French	34.49	72.94	32.48	55.52	33.84	65.92	33.86	68.38
Gujarati	32.59	63.54	29.84	51.81	32.08	64.33	31.39	65.60
Hausa	37.30	67.47	35.16	48.26	35.64	59.22	36.38	59.69
Hindi	37.45	74.20	35.42	54.61	38.14	68.47	37.82	71.51
Igbo	35.44	64.36	32.92	42.32	33.55	62.96	34.42	56.86
Indonesian	35.26	79.06	34.22	60.99	35.99	71.20	36.53	73.66
Japanese	41.30	78.60	38.81	55.37	40.41	78.43	41.00	78.67
Kirundi	33.63	70.66	30.12	49.06	32.27	63.90	32.32	67.49
Korean	39.83	78.51	36.86	62.19	38.98	74.68	38.67	79.27
Kyrgyz	26.54	67.99	24.95	54.10	24.49	69.23	25.36	75.01
Marathi	31.31	67.86	29.28	49.31	27.30	66.20	27.06	66.32
Nepali	38.67	80.57	35.11	62.41	37.62	78.49	38.66	78.44
Oromo	29.61	76.35	26.97	58.98	28.19	78.00	28.95	71.41
Pashto	39.88	76.34	36.91	55.39	39.35	72.90	38.89	74.32
Persian	37.86	79.22	36.77	62.97	37.97	75.01	38.50	77.54
Pidgin	34.41	72.03	33.68	54.00	33.52	69.38	33.63	65.81
Portuguese	33.90	64.24	32.79	40.26	34.40	52.93	35.63	57.67
Punjabi	36.39	64.61	32.56	48.98	35.16	60.76	35.88	63.86
Russian	28.71	69.20	28.28	52.04	29.28	65.32	30.42	65.05
Scottish Gaelic	32.95	68.14	28.55	43.12	31.83	61.42	30.92	58.67
Serbian (Cyrillic)	29.20	62.97	26.72	44.89	27.10	62.19	28.33	62.36
Serbian (Latin)	28.21	59.93	24.85	38.66	25.63	58.28	27.14	58.69
Sinhala	36.04	74.76	31.42	57.27	31.52	68.35	34.36	76.12
Somali	31.35	69.42	29.10	51.16	30.71	62.08	30.17	63.61
Spanish	27.33	62.34	26.90	43.45	27.44	55.76	27.53	59.26
Swahili	36.46	73.25	33.68	54.17	34.98	69.04	34.38	66.90
Tamil	33.26	81.91	31.37	63.84	31.68	77.17	31.25	80.79
Telugu	29.11	67.83	26.39	55.12	27.54	60.45	27.82	68.07
Thai	30.66	73.27	28.43	49.35	29.62	63.97	29.80	71.96
Tigrinya	36.37	76.71	32.24	59.20	34.69	69.19	35.23	73.26
Turkish	33.65	74.66	32.81	51.91	33.01	66.81	33.58	68.67
Ukrainian	28.99	70.87	28.32	54.65	29.08	64.65	29.64	65.28
Urdu	38.18	74.81	36.64	54.70	38.65	70.80	38.67	72.61
Uzbek	28.06	70.62	25.96	52.32	25.40	68.34	26.15	72.99
Vietnamese	36.69	73.01	32.77	53.04	36.07	68.67	37.41	69.45
Welsh	33.89	76.06	31.97	49.66	34.43	66.18	33.93	67.17
Yoruba	37.01	68.48	33.90	50.81	35.03	67.52	35.28	68.20
Average	34.00	72.17	31.85	53.41	33.12	67.39	33.44	69.62

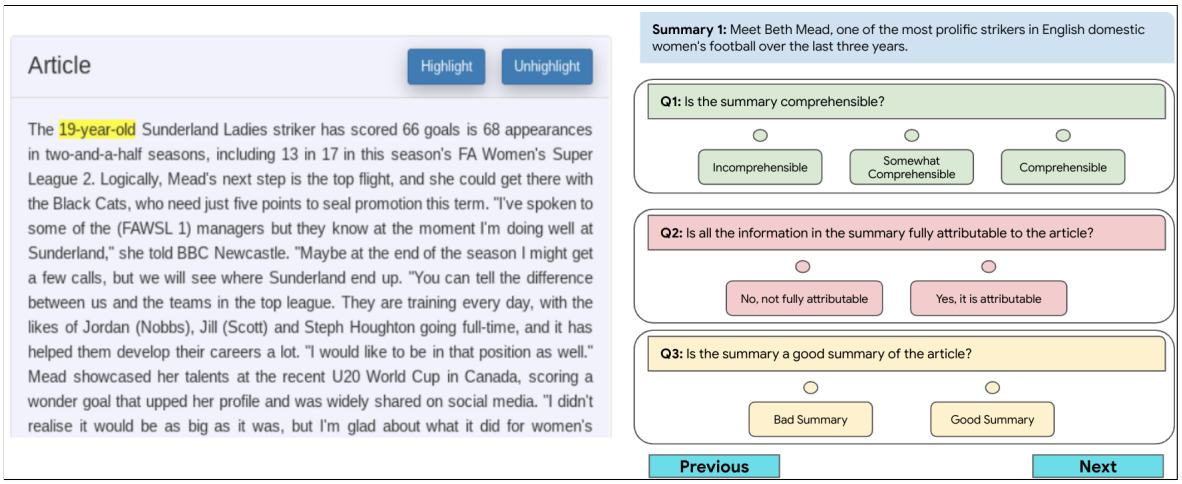
Table 11: ROUGE-L and NLI scores per language on the XLSum test set for our Filtered model vs. comparison systems. For simplicity, all models are compared using their Best-ROUGE checkpoints. XLSum mT5-base predictions are taken from the original XLSum paper (Hasan et al., 2021). However, we report on the recomputed ROUGE-L using the SentencePiece tokenization of mT5 to make it comparable with others. See Section 5.2 for more details on Self-Rouge and Random baselines.

Language	Family	XNLI	Resource
Amharic	Semitic	non-xnli	Low
Arabic	Semitic	xnli	High
Azerbaijani	Turkic	non-xnli	Medium
Bengali	Indo-European	non-xnli	Medium
Burmese	Sino-Tibetan	non-xnli	Low
Chinese Simplified	Sino-Tibetan	xnli	High
Chinese Traditional	Sino-Tibetan	xnli	High
English	Indo-European	xnli	Very High
French	Romance	xnli	Medium
Gujarati	Indo-European	non-xnli	Medium
Hausa	Afro-Asiatic	non-xnli	Medium
Hindi	Indo-European	xnli	Very High
Igbo	Niger-Congo	non-xnli	Low
Indonesian	Austronesian	non-xnli	High
Japanese	Japonic	non-xnli	Medium
Kirundi	Bantu	non-xnli	Low
Korean	Koreanic	non-xnli	Low
Kyrgyz	Turkic	non-xnli	Low
Marathi	Indo-Aryan	non-xnli	High
Nepali	Indo-Aryan	non-xnli	Low
Oromo	Afro-Asiatic	non-xnli	Medium
Pashto	Indo-Iranian	non-xnli	High
Persian	Indo-Iranian	non-xnli	High
Pidgin	Unknown	non-xnli	Medium
Portuguese	Romance	non-xnli	High
Punjabi	Indo-Iranian	non-xnli	Medium
Russian	Indo-European	xnli	High
Scottish Gaelic	Celtic	non-xnli	Low
Serbian Cyrillic	Indo-European	non-xnli	Medium
Serbian Latin	Indo-European	non-xnli	Medium
Sinhala	Indo-European	non-xnli	Low
Somali	Afro-Asiatic	non-xnli	Low
Spanish	Romance	xnli	High
Swahili	Bantu	xnli	Medium
Tamil	Dravidian	non-xnli	High
Telugu	Dravidian	non-xnli	High
Thai	Kra-Dai Languages	xnli	Medium
Tigrinya	Semitic	non-xnli	Low
Turkish	Turkic	xnli	High
Ukrainian	Slavic	non-xnli	High
Urdu	Indo-european	xnli	High
Uzbek	Turkic	non-xnli	Low
Vietnamese	Austroasiatic	xnli	High
Welsh	Celtic	non-xnli	Medium
Yoruba	Niger-Congo	non-xnli	Medium

Table 12: Classification of XLSum languages into families and their membership in XNLI.

	Vanilla				Filtered				Controlled			
	Best-ROUGE		Best-NLI		Best-ROUGE		Best-NLI		Best-ROUGE		Best-NLI	
Language	ROUGE	NLI	ROUGE	NLI	ROUGE	NLI	ROUGE	NLI	ROUGE	NLI	ROUGE	NLI
English	27.94	55.52	28.52	70.70	28.69	72.83	28.42	**80.48**	**28.79**	69.61	28.49	74.58
Varying Number of Training Resource												
High	30.33	48.86	30.53	55.10	**31.50**	63.90	30.87	**66.72**	30.54	57.81	30.64	61.39
Medium	31.51	53.67	30.75	56.36	**31.67**	63.96	31.03	**67.03**	31.31	58.10	30.29	59.94
Low	33.22	56.09	32.20	58.64	**32.84**	63.86	32.04	**70.16**	32.70	57.80	32.09	61.06
Language Families												
Indo-European	33.13	46.37	34.92	47.77	36.67	51.84	38.00	**56.02**	**39.20**	50.31	36.13	51.36
Romance	29.02	42.96	28.47	47.11	**29.36**	55.71	28.79	**61.39**	28.69	51.73	28.61	53.27
Turkic	26.27	54.41	**26.32**	54.09	26.28	61.48	26.09	**68.71**	25.79	57.62	25.85	58.97
Semitic	32.78	51.89	32.52	60.70	**33.38**	65.12	32.21	**70.68**	32.36	55.68	32.79	61.88
Afro-Asiatic	**31.63**	53.43	30.05	58.37	31.17	62.55	30.30	**69.09**	31.15	57.13	30.43	58.86

Table 13: ROUGE-L and NLI scores on XLSum development set for best checkpoints averaged across language groups. For training resources we consider three groups with varying numbers of training examples: High ([70K–10K]), Medium ([10K–6K]), and Low (less than 6K). For language families, the Indo-European cluster represents Bengali, Gujarati, Hindi, russian, Serbian (Cyrillic and Latin), and Sinhala; the Romance cluster comprises of French, Protuguese, and Spanish; the Turkic cluster contains Azerbaijani, Kyrgyz, Turkish, and Uzpek; Semitic languages are Amharic, Arabic, and Tigrinya; the Afro-Asiatic cluster groups together Hausa, Oromo, and Somali; finally, the Indo-Iranian cluster represents Pashto, Persian, and Punjabi; we omit clusters with two members and singletons. We also create two subsets depending on whether they appear in the XNLI dataset used to train our multilingual NLI model (Available) or not (Unavailable). Highest scores are in **bold**.



Q1: Is the summary comprehensible?	
Incomprehensible:	The summary is difficult to understand. It can have serious grammatical errors, low fluency, and/or repeated information.
Somewhat Comprehensible:	The summary generally makes sense but suffers from grammatical errors, low fluency, and/or repeated information.
Comprehensible:	The summary is understandable. It does not exhibit any grammatical errors, disfluencies, and/or repeated information.

Q2: Is all the information in the summary fully attributable to the article?	
Yes, it is attributable:	Select this option if it is accurate to say, "The provided news article says…" or "According to the news article…" with the summary following this phrase.
No, not fully attributable:	Select this option if only some of the information is supported in the news article, but other parts of the information are missing from the news article or not an accurate representation.

Q3: Is the summary a good summary of the article?	
Bad summary:	The summary does not capture the important information in the article, or the captured information is not accurate with the article. It can also exhibit grammatical issues, low fluency, and/or repeated information.
Good Summary:	The summary captures the important information in the article and presents it accurately and concisely. It does not exhibit any grammatical errors, disfluencies, and/or repeated information.

Figure 3: A snapshot of the interface and instructions were used in our human evaluation studies.

Country of Residence	Total Workers	%
Ethiopia	24	6.19
Saudi Arabia	6	1.55
Turkey	15	3.87
Azerbaijan	12	3.09
India	57	14.69
Indonesia	9	2.32
United Kingdom	14	3.61
Argentina	1	0.26
United States	28	7.22
Spain	6	1.55
Pakistan	26	6.70
Czech Republic	1	0.26
France	6	1.55
Nigeria	33	8.51
Japan	7	1.80
South Korea	6	1.55
Kyrgyzstan	8	2.06
Hungary	1	0.26
Myanmar	7	1.80
Nepal	7	1.80
Portugal	8	2.06
Kenya	16	4.12
Burundi	11	2.84
Rwanda	3	0.77
Ukraine	16	4.12
Sri Lanka	6	1.55
Somalia	5	1.29
Serbia	14	3.61
Thailand	5	1.29
Uzbekistan	9	2.32
Vietnam	8	2.06
China	5	1.29
Taiwan	8	2.06
Total	388	100.00

Table 14: Geographic characteristics of our participants.

	Vanilla		Filtered		Controlled		Reference
Language	**Best-ROUGE**	**Best-NLI**	**Best-ROUGE**	**Best-NLI**	**Best-ROUGE**	**Best-NLI**	
amharic	0.74	0.77	0.75	0.7	0.75	0.78	0.75
arabic	0.89	0.91	0.92	0.96	0.93	0.95	0.97
azerbaijani	0.69	0.67	0.68	0.7	0.64	0.68	0.68
bengali	0.92	0.92	0.93	0.94	0.94	0.94	0.94
burmese	0.94	0.91	0.93	0.93	0.91	0.91	0.94
chinese (simp.)	0.85	0.78	0.92	0.88	0.81	0.87	0.95
chinese (trad.)	0.83	0.81	0.86	0.83	0.82	0.86	0.94
english	0.91	0.92	0.91	0.92	0.89	0.91	0.9
french	0.92	0.91	0.91	0.92	0.88	0.94	0.93
gujarati	0.88	0.89	0.85	0.9	0.86	0.9	0.87
hausa	0.98	0.97	0.97	0.97	0.96	0.95	0.99
hindi	0.78	0.79	0.82	0.83	0.74	0.79	0.87
igbo	0.86	0.88	0.85	0.89	0.87	0.88	0.88
indonesian	0.87	0.84	0.88	0.9	0.88	0.92	0.9
japanese	0.85	0.9	0.88	0.85	0.91	0.88	0.94
kirundi	0.79	0.81	0.78	0.82	0.8	0.8	0.88
korean	0.88	0.83	0.9	0.87	0.84	0.88	0.94
kyrgyz	0.93	0.94	0.94	0.96	0.94	0.94	0.96
marathi	0.78	0.76	0.81	0.81	0.81	0.75	0.79
nepali	0.87	0.95	0.92	0.95	0.85	0.94	0.87
oromo	0.61	0.56	0.62	0.61	0.66	0.55	0.78
pashto	0.71	0.7	0.67	0.7	0.7	0.68	0.6
persian	0.81	0.83	0.86	0.85	0.78	0.84	0.84
pidgin	0.87	0.84	0.84	0.84	0.87	0.84	0.84
portuguese	0.92	0.94	0.97	0.95	0.94	0.96	0.97
punjabi	0.71	0.71	0.69	0.74	0.75	0.71	0.67
russian	0.63	0.66	0.64	0.7	0.68	0.68	0.69
scottish gaelic	0.84	0.83	0.83	0.84	0.81	0.84	0.85
serbian (cyrillic)	0.84	0.91	0.81	0.91	0.86	0.88	0.95
serbian (latin)	0.85	0.82	0.8	0.81	0.88	0.84	0.94
sinhala	0.95	0.93	0.97	0.96	0.93	0.94	0.99
somali	0.86	0.84	0.83	0.91	0.86	0.86	0.91
spanish	0.94	0.96	0.95	0.98	0.9	0.98	0.98
swahili	0.93	0.95	0.95	0.95	0.94	0.97	0.92
tamil	0.86	0.84	0.85	0.86	0.84	0.84	0.82
telugu	0.87	0.89	0.87	0.9	0.85	0.9	0.91
thai	0.88	0.87	0.89	0.89	0.87	0.91	0.97
tigrinya	0.88	0.84	0.87	0.88	0.86	0.84	0.96
turkish	0.86	0.85	0.88	0.91	0.86	0.86	0.95
ukrainian	0.9	0.93	0.91	0.94	0.87	0.94	0.97
urdu	0.64	0.67	0.63	0.67	0.68	0.67	0.59
uzbek	0.89	0.83	0.85	0.88	0.83	0.87	0.81
vietnamese	0.97	0.96	0.95	0.95	0.92	0.96	0. 96
welsh	0.88	0.86	0.91	0.87	0.89	0.9	0.96
yoruba	0.83	0.78	0.78	0.78	0.75	0.75	0.85
Avgerage	0.85	0.84	0.85	0.86	0.84	0.86	0.88

Table 15: Mean human judgments for Summary Quality per language on the XLSum *test set* for Best-ROUGE and Best-NLI checkpoints. We also include judgments for Reference summaries.

Language	Vanilla		Filtered		Controlled		Reference
	Best-ROUGE	Best-NLI	Best-ROUGE	Best-NLI	Best-ROUGE	Best-NLI	
amharic	0.47	0.53	0.53	0.5	0.55	0.54	0.39
arabic	0.2	0.29	0.36	0.37	0.34	0.34	0.14
azerbaijani	0.39	0.44	0.46	0.46	0.41	0.41	0.28
bengali	0.79	0.8	0.8	0.84	0.79	0.79	0.61
burmese	0.32	0.39	0.3	0.4	0.33	0.34	0.26
chinese (simp.)	0.47	0.5	0.49	0.62	0.47	0.47	0.31
chinese (trad.)	0.46	0.53	0.54	0.53	0.5	0.5	0.33
english	0.38	0.45	0.46	0.52	0.46	0.4	0.3
french	0.55	0.53	0.58	0.65	0.64	0.57	0.3
gujarati	0.48	0.49	0.51	0.5	0.51	0.47	0.45
hausa	0.29	0.32	0.32	0.35	0.31	0.35	0.15
hindi	0.44	0.5	0.49	0.54	0.48	0.49	0.37
igbo	0.41	0.47	0.4	0.47	0.36	0.44	0.22
indonesian	0.38	0.42	0.43	0.49	0.46	0.43	0.14
japanese	0.11	0.16	0.13	0.19	0.16	0.16	0.07
kirundi	0.31	0.28	0.31	0.41	0.36	0.33	0.23
korean	0.43	0.51	0.48	0.58	0.49	0.44	0.2
kyrgyz	0.59	0.63	0.63	0.67	0.62	0.67	0.34
marathi	0.61	0.65	0.7	0.72	0.68	0.64	0.52
nepali	0.51	0.5	0.45	0.59	0.55	0.48	0.24
oromo	0.46	0.45	0.46	0.5	0.48	0.44	0.48
pashto	0.59	0.6	0.56	0.56	0.61	0.55	0.37
persian	0.26	0.29	0.25	0.26	0.31	0.29	0.12
pidgin	0.29	0.32	0.32	0.41	0.38	0.31	0.16
portuguese	0.29	0.35	0.3	0.42	0.3	0.35	0.17
punjabi	0.49	0.52	0.5	0.58	0.55	0.53	0.31
russian	0.36	0.4	0.38	0.47	0.47	0.44	0.28
scottish gaelic	0.49	0.55	0.52	0.56	0.57	0.56	0.53
serbian (cyrillic)	0.41	0.47	0.43	0.48	0.52	0.45	0.3
serbian (latin)	0.36	0.33	0.33	0.44	0.46	0.36	0.28
sinhala	0.39	0.43	0.31	0.51	0.41	0.38	0.17
somali	0.5	0.56	0.58	0.63	0.52	0.56	0.41
spanish	0.42	0.54	0.52	0.61	0.56	0.51	0.35
swahili	0.52	0.6	0.58	0.66	0.64	0.59	0.39
tamil	0.55	0.55	0.54	0.61	0.57	0.55	0.24
telugu	0.33	0.36	0.35	0.37	0.38	0.36	0.26
thai	0.43	0.46	0.37	0.55	0.5	0.52	0.3
tigrinya	0.5	0.56	0.53	0.65	0.62	0.5	0.34
turkish	0.52	0.55	0.61	0.62	0.57	0.6	0.47
ukrainian	0.54	0.56	0.58	0.67	0.6	0.57	0.39
urdu	0.56	0.61	0.52	0.61	0.6	0.6	0.41
uzbek	0.63	0.55	0.58	0.64	0.57	0.6	0.4
vietnamese	0.44	0.43	0.44	0.54	0.47	0.45	0.21
welsh	0.36	0.3	0.41	0.44	0.37	0.41	0.25
yoruba	0.42	0.38	0.42	0.38	0.41	0.39	0.33
Average	0.44	0.47	0.46	0.52	0.49	0.47	0.31

Table 16: Mean human judgments for Attribution per language on the XLSum test set for Best-ROUGE and Best-NLI checkpoints. We also include judgments for Reference summaries.

	Vanilla		Filtered		Controlled		Reference
Language	**Best-ROUGE**	**Best-NLI**	**Best-ROUGE**	**Best-NLI**	**Best-ROUGE**	**Best-NLI**	
amharic	0.37	0.38	0.4	0.36	0.43	0.44	0.32
arabic	0.16	0.24	0.29	0.33	0.28	0.3	0.14
azerbaijani	0.36	0.43	0.42	0.45	0.39	0.38	0.28
bengali	0.77	0.76	0.78	0.82	0.76	0.77	0.61
burmese	0.29	0.35	0.27	0.35	0.28	0.32	0.23
chinese (simp.)	0.25	0.24	0.27	0.34	0.24	0.27	0.19
chinese (trad.)	0.4	0.44	0.48	0.43	0.45	0.41	0.32
english	0.33	0.39	0.41	0.45	0.38	0.33	0.27
french	0.35	0.38	0.39	0.43	0.43	0.38	0.23
gujarati	0.32	0.34	0.33	0.33	0.35	0.36	0.28
hausa	0.28	0.32	0.29	0.32	0.3	0.32	0.15
hindi	0.39	0.45	0.42	0.51	0.42	0.44	0.34
igbo	0.35	0.4	0.34	0.42	0.32	0.38	0.19
indonesian	0.33	0.37	0.4	0.45	0.39	0.4	0.13
japanese	0.11	0.13	0.11	0.13	0.12	0.11	0.07
kirundi	0.27	0.25	0.26	0.35	0.34	0.29	0.21
korean	0.34	0.36	0.39	0.42	0.36	0.33	0.17
kyrgyz	0.53	0.59	0.6	0.63	0.59	0.64	0.31
marathi	0.57	0.59	0.65	0.66	0.64	0.57	0.5
nepali	0.35	0.37	0.35	0.47	0.39	0.37	0.17
oromo	0.28	0.26	0.27	0.33	0.35	0.23	0.36
pashto	0.39	0.41	0.38	0.4	0.37	0.39	0.26
persian	0.29	0.31	0.29	0.29	0.29	0.33	0.15
pidgin	0.26	0.29	0.29	0.37	0.35	0.28	0.16
portuguese	0.28	0.33	0.3	0.4	0.26	0.33	0.17
punjabi	0.48	0.51	0.49	0.56	0.52	0.52	0.31
russian	0.32	0.38	0.34	0.43	0.39	0.4	0.27
scottish gaelic	0.41	0.46	0.4	0.42	0.46	0.42	0.4
serbian (cyrillic)	0.34	0.43	0.34	0.41	0.43	0.38	0.28
serbian (latin)	0.31	0.28	0.23	0.36	0.41	0.3	0.26
sinhala	0.35	0.37	0.28	0.47	0.37	0.35	0.16
somali	0.46	0.52	0.54	0.61	0.51	0.51	0.42
spanish	0.25	0.34	0.32	0.39	0.33	0.32	0.23
swahili	0.46	0.56	0.53	0.61	0.58	0.53	0.35
tamil	0.54	0.55	0.54	0.61	0.55	0.53	0.24
telugu	0.32	0.33	0.35	0.37	0.36	0.35	0.29
thai	0.38	0.39	0.34	0.49	0.46	0.47	0.29
tigrinya	0.44	0.46	0.46	0.56	0.54	0.43	0.32
turkish	0.49	0.52	0.55	0.57	0.54	0.57	0.46
ukrainian	0.47	0.49	0.52	0.63	0.49	0.54	0.37
urdu	0.46	0.48	0.4	0.46	0.51	0.48	0.38
uzbek	0.56	0.48	0.49	0.58	0.47	0.53	0.34
vietnamese	0.38	0.34	0.38	0.45	0.38	0.37	0.19
welsh	0.29	0.24	0.36	0.34	0.29	0.37	0.23
yoruba	0.36	0.28	0.35	0.29	0.33	0.31	0.27
Average	0.37	0.40	0.39	0.45	0.41	0.40	0.27

Table 17: Mean human judgments for Informativeness per language on the XLSum test set for Best-ROUGE and Best-NLI checkpoints. We also include judgments for Reference summaries.

Language	Vanilla		Filtered		Controlled		Reference
	Best-ROUGE	Best-NLI	Best-ROUGE	Best-NLI	Best-ROUGE	Best-NLI	
English	0.91	**0.92**	0.91	**0.92**	0.89	0.91	0.9
Training Resources							
High	0.83	0.83	0.85	0.86	0.83	0.85	**0.87**
Medium	0.84	0.84	0.83	0.85	0.84	0.84	**0.88**
Low	0.87	0.86	0.87	0.88	0.86	0.87	**0.89**
Language Family							
Indo-European	0.81	0.82	0.81	0.84	0.82	0.83	**0.86**
Romance	0.93	0.94	0.94	0.95	0.9	**0.96**	**0.96**
Turkic	0.84	0.82	0.84	**0.86**	0.82	0.84	0.85
Semitic	0.83	0.84	0.84	0.85	0.85	0.86	**0.90**
Afro-Asiatic	0.82	0.79	0.81	0.83	0.83	0.78	**0.89**
Indo-Iranian	0.74	0.75	0.74	**0.76**	0.74	0.74	0.70
XNLI Training Data							
Available	0.85	0.85	0.86	0.88	0.84	0.87	**0.89**
Unavailable	0.84	0.84	0.84	0.86	0.84	0.85	**0.87**

Table 18: Mean human judgments on Summary Quality for the best checkpoints averaged across language groups with 1) varying number of training resources, 2) language families and 3) depending on whether XNLI data is available. See Table 4 for more details about different groups. Best results in each row are in **bold**.

Language	Vanilla		Filtered		Controlled		Reference
	Best-ROUGE	Best-NLI	Best-ROUGE	Best-NLI	Best-ROUGE	Best-NLI	
English	0.38	0.45	0.46	**0.52**	0.46	0.4	0.3
Training Resources							
High	0.44	0.48	0.47	**0.53**	0.49	0.48	0.30
Medium	0.42	0.44	0.44	**0.49**	0.48	0.45	0.31
Low	0.46	0.49	0.47	**0.55**	0.50	0.49	0.31
Language Family							
Indo-European	0.48	0.5	0.47	**0.55**	0.53	0.5	0.36
Romance	0.42	0.47	0.47	**0.56**	0.5	0.48	0.28
Turkic	0.53	0.54	0.57	**0.60**	0.54	0.57	0.37
Semitic	0.39	0.46	0.47	**0.51**	0.50	0.46	0.29
Afro-Asiatic	0.42	0.44	0.45	**0.49**	0.43	0.45	0.35
Indo-Iranian	0.45	0.47	0.44	0.47	**0.49**	0.46	0.27
XNLI Training Data							
Available	0.44	0.49	0.49	**0.56**	0.52	0.50	0.32
Unavailable	0.44	0.46	0.45	**0.51**	0.48	0.46	0.30

Table 19: Human evaluation results for Attribution for the best checkpoints averaged across language groups with 1) varying number of training resources, 2) language families and 3) depending on whether XNLI is available. See Table 4 for more details about different groups. Best results in each row are in **bold**.

Language	Vanilla		Filtered		Controlled		Reference
	Best-ROUGE	Best-NLI	Best-ROUGE	Best-NLI	Best-ROUGE	Best-NLI	
English	0.33	0.39	0.41	**0.45**	0.38	0.33	0.27
Training Resources							
High	0.37	0.40	0.40	**0.45**	0.40	0.41	0.27
Medium	0.36	0.37	0.37	**0.41**	0.40	0.38	0.28
Low	0.39	0.42	0.40	**0.47**	0.42	0.42	0.27
Language Family							
Indo-European	0.41	0.44	0.39	**0.47**	0.45	0.43	0.32
Romance	0.3	0.35	0.34	**0.41**	0.34	0.34	0.21
Turkic	0.48	0.5	0.52	**0.56**	0.50	0.53	0.35
Semitic	0.32	0.36	0.39	**0.41**	**0.41**	0.39	0.26
Afro-Asiatic	0.34	0.37	0.37	**0.42**	0.39	0.35	0.31
Indo-Iranian	0.38	0.41	0.39	**0.42**	0.40	0.41	0.24
XNLI Training Data							
Available	0.36	0.40	0.39	**0.45**	0.41	0.41	0.28
Unavailable	0.38	0.39	0.39	**0.44**	0.41	0.40	0.27

Table 20: Human evaluation results for Informativeness for the best checkpoints averaged across language groups with 1) varying number of training resources, 2) language families and 3) depending on whether XNLI data is available. See Table 4 for more details about different groups. Best results in each row are in **bold**.

Document
Aucun cas de choléra n'avait été détecté en Algérie depuis 1996, tandis que la dernière épidémie d'ampleur remonte à 1986. Le précédent bilan de l'actuel épisode, communiqué vendredi, faisait état de 41 cas avérés, dont un décès, pour 88 cas suspects, répartis à Alger -5 malades confirmés- et dans trois wilayas (préfectures) alentour. Un deuxième patient est décédé du choléra dans la wilaya de Blida (50 km au sud d'Alger), où 25 cas sont confirmés, a annoncé samedi le ministère. Selon lui, un cas a été confirmé dans la wilaya de Médéa et un cas suspect recensé dans celle d'Aïn Defla, toutes deux jusqu'ici épargnées. Situées respectivement à 80 et 100 km au sud-ouest d'Alger, ces deux régions sont limitrophes de celles de Blida et Tipaza (70 km à l'ouest d'Alger), où la grande majorité des malades sont recensés.

Reference Summary
Selon les dernières analyses, "46 cas de choléra ont été confirmés parmi les 139 cas hospitalisés depuis le 7 août 2018", a fait savoir le ministère dans un communiqué. *According to the latest analyses, "46 cases of cholera have been confirmed among the 139 cases hospitalized since August 7, 2018," the ministry said in a statement.*

Vanilla
Le ministère de la santé algérien a annoncé samedi que l'épidémie de choléra a fait plusieurs morts et que le nombre de cas suspects a atteint les 100.. *The Algerian Ministry of Health announced on Saturday that the cholera epidemic has caused several deaths and that the number of suspected cases has reached 100..*

Controlled
Le ministère de la Santé a confirmé samedi l'augmentation du nombre de cas de choléra en Algérie, dont deux morts. *The Ministry of Health confirmed on Saturday the increase in the number of cholera cases in Algeria, including two deaths.*

Filtered
Le ministère algérien de la Santé a annoncé samedi que 41 cas de choléra ont été confirmés et 88 cas suspects, selon le ministère. *The Algerian Ministry of Health announced on Saturday that 41 cases of cholera have been confirmed and 88 suspected cases, according to the ministry.*

Table 21: Input XLSum document in French, accompanied by reference summary, and summaries generated by the Vanilla, Controlled, and Filtered models, respectively. English translations of the summaries are shown in *italics*.

Document
En septembre dernier, les essais d'un autre vaccin anti-Ebola avaient été lancés, également à Oxford. Les derniers essais vont porter sur 72 volontaires dont l'âge varie de 18 à 50 ans. Des tests préliminaires sur des singes avaient montré que le vaccin, mis au point par Janssen Pharmaceutical Companies, confère une immunité contre Ebola. Les volontaires d'Oxford sont les premiers humains à se soumettre au test de ce vaccin expérimental. Dr Matthew Snape, de l'Oxford Vaccine Group, une cellule du département de Pédiatrie de l'Université d'Oxford, déclare: "notre objectif est d'immuniser tous les participants au bout d'un mois." "Il s'agit principalement de chercher à comprendre le profil du vaccin en matière de sécurité." Au cours de ces essais, les volontaires recevront des doses supplémentaires un ou deux mois après l'injection de la dose initiale. Des essais similaires, à petite échelle, sont en cours aux Etats-Unis et dans trois autres pays africains épargnés par l'épidémie. La dose initiale est destinée à préparer le système immunitaire du sujet, et la deuxième vient doper l'orgacd /home/mlap/google/face/latex/nisme à réagir en cas d'infection.

Reference Summary
Des chercheurs à l'Université d'Oxford ont commencé l'expérimentation d'un vaccin contre Ebola sur des volontaires bien portants. *Researchers at the University of Oxford have begun experimenting with a vaccine against Ebola on healthy volunteers.*

Vanilla
Des essais à grande échelle d'un vaccin contre Ebola sont en cours à Oxford, dans le sud de l'Angleterre. *Large-scale trials of an Ebola vaccine are underway in Oxford, southern England.*

Controlled
Les essais d'un vaccin anti-Ebola sont en cours à l'Université d'Oxford. *Trials of an Ebola vaccine are underway at the University of Oxford.*

Filtered
Les scientifiques de l'Université d'Oxford ont lancé les essais d'un vaccin expérimental contre Ebola. *Scientists at the University of Oxford have launched trials of an experimental Ebola vaccine.*

Table 22: Input XLSum document in French, accompanied by reference summary, and summaries generated by the Vanilla, Controlled, and Filtered models, respectively. English translations of the summaries are shown in *italics*.

Document
Los ecologistas están acusados de piratería. En la imagen, El portavoz de Greenpreace Internacional, Roman Dolgov. Los argentinos Miguel Hernan Pérez Orz y Camila Speziale, junto con la brasileña Ana Paula Alminhana Maciel, se enfrentan a cargos de piratería, lo que puede ser penado con hasta 15 años de cárcel en Rusia. El presidente, Vladimir Putin, señaló que los acusados no son piratas, pero no se opuso a las detenciones. Los miembros de Greenpeace arrestados provienen de 18 países. Un tribunal de Rusia ordenó el jueves extender la detención por dos meses a cinco activistas del grupo de arrestados, que presuntamente intentaron escalar una plataforma petrolera en alta mar en el Ártico ruso. Los ecologistas afectados por la medida provienen de Rusia, Polonia, Canadá y Nueva Zelanda. Todos los detenidos esperan ser llamados por el tribunal de la ciudad de Murmansk, en el norte de Rusia, para que cada uno de sus casos sea escuchado. Los miembros de Greenpeace aseguran que organizaron una protesta legal y pacífica.
Reference Summary
Tres latinoamericanos se encuentran en la lista de los 30 activistas de Greenpeace arrestados en Rusia por protestar la semana pasada contra la exploración de petróleo en aguas árticas. *Three Latin Americans are on the list of 30 Greenpeace activists arrested in Russia for protesting last week against oil exploration in Arctic waters.*
Vanilla
Activistas argentinos, brasileños y argentinos, que fueron arrestados en Rusia por protestar en el mar Negro en protesta por el cierre de una plataforma petrolera en el norte de Rusia, están acusados de piratería. *Argentine, Brazilian and Argentinian activists, who were arrested in Russia for protesting in the Black Sea to protest the closure of an oil rig in northern Russia, are accused of piracy.*
Controlled
Un tribunal ruso ordenó la detención de cinco miembros del grupo ecologista Greenpeace, acusados de intentar escalar una plataforma petrolera en el Ártico. *A Russian court ordered the arrest of three members of the environmental group Greenpeace, accused of trying to scale an oil platform in the Arctic.*
Filtered
Los dos argentinos que fueron arrestados en Rusia por protestar contra el desarrollo de una plataforma petrolera en el Ártico están en riesgo de ser encarcelados por piratería. *The two Argentines who were arrested in Russia for protesting against the development of an oil rig in the Arctic are at risk of being jailed for piracy.*

Table 23: Input XLSum document in Spanish, accompanied by reference summary, and summaries generated by the Vanilla, Controlled, and Filtered models, respectively. English translations of the summaries are shown in *italics*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We wrote a limitations section.

☑ A2. Did you discuss any potential risks of your work?
While our work is not introducing new risks, it is relevant to the risk of spreading misinformation with text generation systems due to factual consistency issues, which we discuss in the ethics section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
The abstract and introduction sections include all the main claims and contributions of the paper.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We used the XLSum dataset throughout the paper, and cited its authors throughout the paper (Hasan et al 2021).

☑ B1. Did you cite the creators of artifacts you used?
We used the XLSum dataset throughout the paper, and cited its authors throughout the paper (Hasan et al 2021).

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The dataset is available for non-commercial use: https://github.com/csebuetnlp/xl-sum#license

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Mentioned https://github.com/csebuetnlp/xl-sum#license in section 1

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We did not collect new data, but only annotated existing data that was collected from the publicly available BBC website.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We described all the languages that are part of the XLSum dataset in Table 1.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
We described the metadata in Table 1.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We reported the number of parameters in section 5.1 and the compute infrastructure details in Appendix B.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We report those details in appendix B.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We report results based on a single run and mention this in section 6.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We added a reference to the ROUGE implementation we used and other relevant details in section 5.3.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 5.4

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We show the instructions and screenshots in the appendix.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We provided those details in appendix D.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
We only used the XLSum data for annotations which is allowed for research purposes: https://github.com/csebuetnlp/xl-sum#license

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We did not collect new data but only annotated existing data which is open for research purposes.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We provided those details in appendix D.