CCSum: A Large-Scale and High-Quality Dataset for Abstractive News Summarization

Xiang Jiang
Amazon
jxiang@amazon.com

Markus Dreyer
Amazon
mddreyer@amazon.com

Abstract

Training a supervised news summarization model requires large amounts of high-quality training data consisting of news articles paired with reference summaries. However, obtaining such data is costly, and existing datasets contain considerable amount of noise. We present a new large-scale and high-quality dataset for supervised abstractive news summarization containing 1.3 million training samples, which we call CCSUM.[1] In creating this dataset, we take advantage of the journalistic inverted-pyramid style in news writing: In some articles, the first sentence can be considered a summary of the reported story. Accordingly, among 35 million CommonCrawl News articles, we identify pairs of articles about the same news story and use one article's first sentence as the summary for the other article. To ensure high quality, we apply strict filters whose parameters we optimize using Bayesian optimization. We show that the resulting dataset is more factual and informative than established summarization datasets; less than 1% of the summaries have major factual inconsistencies with the corresponding news articles, compared to 5.5% to 15.4% in existing datasets, according to our human evaluation. Summarization models trained on our dataset are more favored compared to those trained on CNN/Daily Mail. The proposed dataset can open new opportunities for future research in abstractive summarization.

1 Introduction

News summarization is the task of automatically converting news articles into concise summaries that correctly convey the main information of the articles. Particularly, in *abstractive* news summarization, summaries are generated as coherent reformulations of the articles' main points, rather than verbatim extraction of the articles' phrases and sentences. Abstractive news summarization has made

---

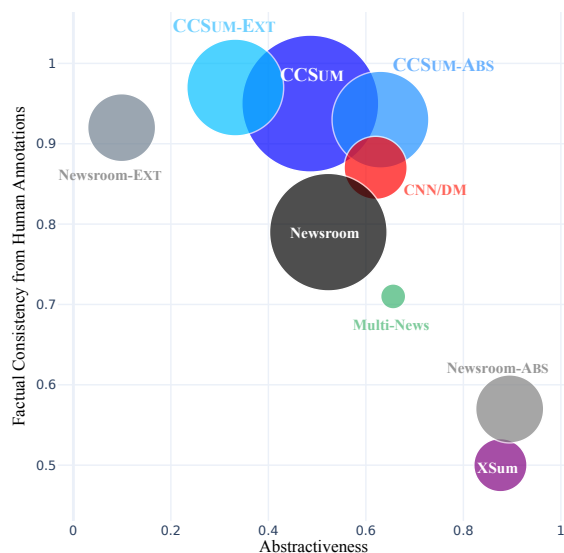[1]CCSUM is available at https://github.com/amazon-science/ccsum.

Figure 1: News summaries in our large-scale dataset CCSUM (variants: CCSUM-EXT and CCSUM-ABS) have higher factual consistency than those of existing datasets, while written in an abstractive style comparable to CNN/DM, Multi-News and Newsroom. Bubble size corresponds to dataset size. (Details in Section 4.)

significant progress since the advent of transformers (Vaswani et al., 2017): Summarization models have grown larger (Raffel et al., 2020), objective functions have evolved (Cao and Wang, 2021) and strategies for more factual (Zhu et al., 2021) or coherent (Gunel et al., 2020) summaries have been presented. However, the datasets on which abstractive news summarizers are trained have remained constant for several years. CNN/Daily Mail (CNN/DM; Hermann et al. (2015)) and XSum (Narayan et al., 2018) are used most frequently, Multi-News (Fabbri et al., 2019) and Newsroom (Grusky et al., 2018) have found adoption as well. While these datasets differ in size (Figure 1), they contain considerable amount of noise (Tejaswin et al., 2021). In human evaluations, summaries generated in zero-shot fashion by large language models have been preferred over sum-
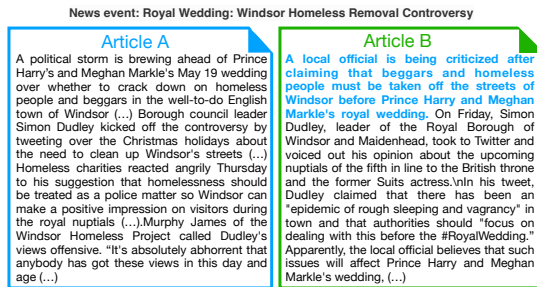
Article A

A political storm is brewing ahead of Prince Harry's and Meghan Markle's May 19 wedding over whether to crack down on homeless people and beggars in the well-to-do English town of Windsor (…) Borough council leader Simon Dudley kicked off the controversy by tweeting over the Christmas holidays about the need to clean up Windsor's streets (…) Homeless charities reacted angrily Thursday to his suggestion that homelessness should be treated as a police matter so Windsor can make a positive impression on visitors during the royal nuptials (…).Murphy James of the Windsor Homeless Project called Dudley's views offensive. "It's absolutely abhorrent that anybody has got these views in this day and age (…)

Article B

A local official is being criticized after claiming that beggars and homeless people must be taken off the streets of Windsor before Prince Harry and Meghan Markle's royal wedding. On Friday, Simon Dudley, leader of the Royal Borough of Windsor and Maidenhead, took to Twitter and voiced out his opinion about the upcoming nuptials of the fifth in line to the British throne and the former Suits actress.\nIn his tweet, Dudley claimed that there has been an "epidemic of rough sleeping and vagrancy" in town and that authorities should "focus on dealing with this before the #RoyalWedding." Apparently, the local official believes that such issues will affect Prince Harry and Meghan Markle's wedding, (…)

Figure 2: Articles covering the same news event, where Article B's first sentence can summarize Article A.

maries generated by supervised models trained on the above datasets (Goyal et al., 2023).

To mitigate these shortcomings, we introduce a new dataset for abstractive news summarization that is both large-scale and of high quality. It contains 1.3 million pairs of articles and summaries. Less than 1% of the summaries have major factual inconsistencies with the corresponding news articles, compared to 5.5% to 15.4% in existing datasets, according to our human evaluation in Table 3. 86.7% and 95.8% of the reference summaries in our dataset are rated "very good" in informativeness and coherence, compared to only 30.0% to 78.3% and 54.1% to 87.8%, respectively. Our human evaluation shows that summaries generated from a model trained on our dataset are preferred over models trained on CNN/DM, reducing the gap to zero-shot summaries from GPT-3.5 (OpenAI, 2023).

Our method is based on the insight that news articles often follow the journalistic principle of the inverted pyramid (Pöttker, 2003): Articles start with the most important information, followed by important details and, finally, background information and other facts. The most fundamental and important information is often contained in the first sentence—the lead sentence (see Figure 2). Accordingly, we pair the lead sentence from one article as summary for another article describing the same news story. In particular, we obtain 35 million news articles from CommonCrawl News and cluster them such that each cluster contains multiple articles that describes the same news story. In each cluster, we consider all pairs of articles; in each pair, we consider one article's lead sentence as the summary of the other article (Figure 3). This results in over 100 million candidate pairs, each one consisting of a full article and an associated lead sentence from a related article as its potential sum-

mary. Only a small percentage of these candidate pairs are of high quality. Therefore, we construct a multi-aspect filter involving automatic metrics of factual consistency, information coverage, abstractiveness, and other dimensions. We optimize this filter using Bayesian optimization and apply it to remove 99% of the candidate pairs, resulting in the final dataset of articles and lead sentence summaries, which we split into a train, dev, and test portion. We call this dataset CCSUM because it is derived from the CommonCrawl corpus. In summary, our contributions are:

1. We propose a method of constructing high-quality abstractive news summaries at scale, which resulted in CCSUM, a dataset of 1.3 million article-summary pairs for single-document abstractive summarization. CommonCrawl is a continuously crawled corpus, so our dataset can be expanded over time.

2. We benchmark CCSUM alongside established abstractive summarization datasets, and find that CCSUM is more factual and informative based on automatic and human evaluations.

3. We carry out summarization experiments on CCSUM to characterize its impact on summarization model training. We find that CCSUM-trained models are more favored compared to those trained with CNN/DM, narrowing the preference gap with GPT-3.5.

Our dataset can offer new opportunities for text generation research; this could include exploring the effects of model and data scaling, improving factual consistency in generated summaries, and developing better instruction-following models.

2 Related Work

CNN/Daily Mail (CNN/DM) (Hermann et al., 2015) is a widely used dataset consisting of news articles from the CNN and Daily Mail websites, paired with bullet-point descriptions. It has been used for summarization (See et al., 2017) by concatenating each description into a multi-sentence summary with 287k training samples. However, since the descriptions were not intended as summaries, some summaries are not coherent or may not accurately reflect the article content. **XSum** (Narayan et al., 2018) is a popular abstractive summarization dataset consisting of BBC news articles; the first sentence of each article is removed
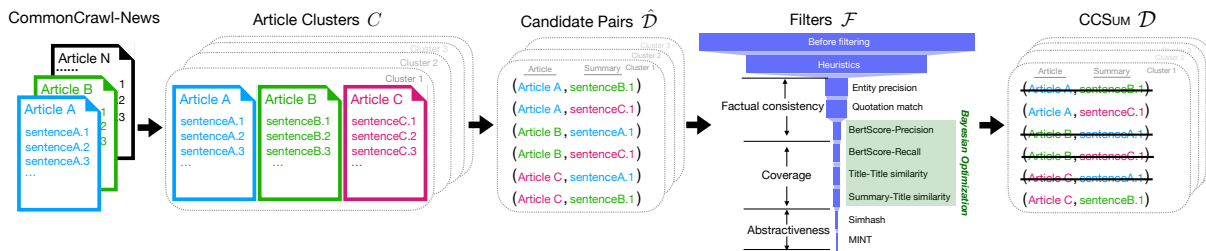
Figure 3: Process to construct the CCSᴜᴍ dataset. News articles from CommonCrawl-News are clustered into news events, from which we generate candidate article-summary pairs, and apply extensive filtering to obtain the final dataset. For instance, out of 312,544 articles published between September 21 and September 23, 2020, 71,731 unique news events were identified, creating 1.8 million candidate pairs. After extensive filtering, 99% of these pairs were eliminated ensuring only high-quality pairs were included in the final dataset.

and used as summary for the remaining part of the article. Its training set has 204k samples. The introductory sentences were not intended to be abstractive summaries and might contain information not present in the article. **Newsroom** (Grusky et al., 2018) was constructed from web-scale crawling and uses the "description" field in each article's HTML metadata as the summary. The summaries are split into extractive, abstractive, and mixed bins, according to the density metric. The summaries tend to be noisy due to the inconsistent quality of HTML metadata. It contains 1.3 million training samples. **Multi-News** (Fabbri et al., 2019) pairs human-written news aggregation summaries with the articles originally linked from those summaries. This makes it a multi-document summarization dataset—in contrast to the other datasets discussed. It contains 44k training samples. **Gigaword** (Rush et al., 2015) pairs articles with their headlines, e.g., "five dead as powerful quake hits southern iran", rather than fully formulated summaries.

3 Dataset Construction

3.1 Overview

We derive the CCSᴜᴍ dataset from 35 million news articles from the Common Crawl News[2] corpus, covering all English-language news articles from 2018 to 2022. Figure 3 illustrates our method for extracting high-quality abstractive summaries. This technique begins by clustering news articles $A_i \in \mathcal{A}$ into news events $\mathcal{C}$. Within each event, we create candidate article-summary pairs by selecting the first sentence from one article A_i^1 as a potential summary for another article A_j. Subsequently, as depicted in Figure 3, we define a set of summary quality filters and use Bayesian optimization to fine-

[2]https://commoncrawl.org/blog/news-dataset-available

tune their parameters. The final CCSᴜᴍ dataset comprises only 1% of these candidate summaries that have successfully met all filter criteria.

3.2 Clustering for News Event Detection

We begin by grouping all news articles into three-day time windows $T_{[t_{\min}, t_{\max})}$ based on their publication times to maintain the temporal coherence of articles within each news event. For articles in each time window $T_{[t_{\min}, t_{\max})}$, we use sentence-BERT (Reimers and Gurevych, 2019) to encode the main text and perform soft clustering (Bezdek, 2013). Note that in soft clustering, an article can be assigned to multiple clusters, which allows us to over-generate candidate article-summary pairs followed-by strict filtering to obtain the final large-scale and high-quality dataset.

3.3 Candidate Summary Generation

We construct candidate summaries $\hat{\mathcal{D}} = \left\{ \left(A_i, A_j^1 \right) \mid i \neq j, A_i \in C, A_j \in C \right\}$ for each cluster C, where the first sentence A_j^1 from article A_j is proposed as an abstractive summary for article A_i. Given the combinatorial nature of this approach, a large number of candidate article-summary pairs are generated— three-day's articles generated over one million candidate summaries.

It should be emphasized that simply having semantic similarity between news articles does not ensure that the first sentence will effectively function as a high-quality abstractive summary of the other article. To address this, we have established a series of filters to ensure that the chosen abstractive summary is relevant, factual, and encapsulates the key points of the article. Figure 3 illustrates the data filtering funnel, highlighting how 99% of the initial article-summary candidates were eliminated to produce the final CCSᴜᴍ dataset.

3.4 Summary Quality Filter Definition

We define a set of metrics, denoted as $\mathcal{F}$, to evaluate the quality of each candidate article-summary pair $\left(A_i, A_j^1\right) \in \hat{\mathcal{D}}$. Each metric f_x is associated with a specific filter constraint, defined by some threshold λ_{f_x}. The final CCSUM dataset $\mathcal{D}$ consists of article-summary pairs from $\hat{\mathcal{D}}$ that successfully meet all filter criteria, where $\mathcal{D} = \left\{ \left(A_i, A_j^1\right) \in \hat{\mathcal{D}} \mid \forall f_x \in \mathcal{F}, f_x\left(A_i, A_j^1\right) > \lambda_{f_x} \right\}$.

Heuristics. We first define a set of heuristics specifically designed to mitigate challenges raised by our candidate summary generation approach. For instance, we ensure the summary ends in proper punctuation and the summary contains at least one entity. More details are described in Appendix D.1.

Factual Consistency. A valid summary must accurately reflect the content of the original news article while avoiding introducing unsupported information. We use the following metrics for factual consistency: (i) *Entity precision* f_{ep} quantifies the proportion of entities in the summary that also appear in the article (Nan et al., 2021). This helps determine if the summary introduces entities absent from the original article. (ii) *BERTScore-Precision* (BS-Precision) f_{BS-P} (Zhang et al., 2019). BertScore Precision works well to evaluate the factual consistency of an abstractive summary against an article (Pagnoni et al., 2021). We calculate BERTScores using both BERT (Devlin et al., 2018) and BART (Lewis et al., 2019; Yuan et al., 2021), as the models complement each other with different pre-training objectives. (iii) *Quotation exact match* f_{quo}: We use regular expression to detect quotations from the summary and ensure their presence in the article.

Coverage. A well-crafted summary should capture the key points of an article while avoiding extraneous details. We use the following metrics to measure coverage: (i) *BERTScore-Recall* (BS-Recall) f_{BS-R} (Zhang et al., 2019) aligns tokens in the article with tokens in the candidate summary and computes the average similarity score over tokens in the article. (ii) *Title-title similarity* f_{t-t} measures the semantic similarity between the article's title and the summary's title, using cosine similarity of sentenceBERT (Reimers and Gurevych, 2019) embeddings. This metric determines if the summary's topic aligns with that of the article. (iii) *Summary-title similarity* f_{s-t} mea-

sures the semantic similarity between the candidate summary A_j^1 and the title of the article A_i.

Abstractiveness. A well-crafted abstractive summary should effectively synthesize and condense the content of an article, rather than merely replicating sentences or phrases from it. We use MINT f_{MINT} (Dreyer et al., 2023) and *Simhash* $f_{simhash}$ (Charikar, 2002; Henzinger, 2006) to measure the abstractiveness of the candidate summary.

3.5 Bayesian Optimization

Given the wide range of filters defined in Section 3.4 and their intricate interactions, manually setting thresholds for each filter is impractical. We use Bayesian Optimization (Kushner, 1964) to address this challenge by performing constrained optimization over filter parameters (Frazier, 2018).

The Search Space. We perform filter parameter search on all the embedding-based metrics, i.e., BERTScore-Precision, BERTScore-Recall, title-title similarity and article-title similarity, which we denote as a vector λ. The remaining filter parameters are set as constant in Table 12 of the Appendix.

Collecting Human Annotations. Bayesian optimization requires a measurable objective function to optimize in the parameter space. We use human judgements on a held-out set of candidate article-summary pairs to guide the search for filter parameters. The intuition is that the ideal filter parameterization should maximize the recall of high-quality summaries while reducing the presence of low-quality summaries. We randomly select 1,000 examples from our candidate dataset $\hat{\mathcal{D}}$ for human annotation. Each candidate article-summary pair is annotated into one of the three categories: no factual error, minor factual error, and major factual error. Please refer to Appendix F.1 for the detailed annotation guidelines. We obtain an annotated dataset $\hat{\mathcal{D}}_{labelled}$ where each example (A_i, A_j^1, y) is associated with a factual consistency label y.

Optimization Objective. The optimization consists of a primary objective and two constraints. The primary objective is the recall of factually correct summaries $f_{recall}(\lambda, \hat{\mathcal{D}}_{labelled})$, where

$$f_{recall} = \frac{\left| \left\{ (A_i, a_j^1, y) \in \mathcal{D}_{labelled} \mid y = \text{correct} \right\} \right|}{\left| \left\{ (A_i, a_j^1, y) \in \hat{\mathcal{D}}_{labelled} \mid y = \text{correct} \right\} \right|}$$

(1)

However, focusing solely on recall could result in a trivial solution, i.e., removing all filters. Therefore,

	Total	Date range	EXT	ABS
Train	1,349,911	1/2018 - 12/2021	674,939	674,972
Val.	10,000	1/2022 - 5/2022	4,853	5,147
Test	10,000	6/2022 - 12/2022	5,053	4,947

Table 1: CCSUM Dataset overview.

we introduce two constraints to ensure that the final dataset is both large-scale and of high quality. The first constraint limits the major factual error rate $f_{\mathrm{error_{major}}}(\lambda, \hat{\mathcal{D}}_{\mathrm{labelled}})$, defined in Eq. 2, after filtering to be less than 3%.

$$f_{\mathrm{error_{major}}} = \frac{|\{(A_i, a_j^1, y) \in D_{\mathrm{labelled}} \mid y = \mathrm{error_{major}}\}|}{|D_{\mathrm{labelled}}|} \tag{2}$$

The second constraint focuses on the precision of factually correct summaries $f_{\mathrm{precision}}(\lambda, \hat{\mathcal{D}}_{\mathrm{labelled}})$, defined in Eq. 3. It requires that the percentage of factually correct summaries exceeds eighty percent after filtering.

$$f_{\mathrm{precision}} = \frac{|\{(A_i, a_j^1, y) \in \mathcal{D}_{\mathrm{labelled}} \mid y = \mathrm{correct}\}|}{|\mathcal{D}_{\mathrm{labelled}}|} \tag{3}$$

Eq. 4 defines the overall optimization objective:

$$\max_{\lambda} \quad f_{\mathrm{recall}}(\lambda, \hat{\mathcal{D}}_{\mathrm{labelled}})$$
$$\text{subject to} \quad f_{\mathrm{error_{major}}}(\lambda, \hat{\mathcal{D}}_{\mathrm{labelled}}) < 0.03, \tag{4}$$
$$f_{\mathrm{precision}}(\lambda, \hat{\mathcal{D}}_{\mathrm{labelled}}) > 0.8$$

Optimization Outcome. On a held-out set of annotated data, the optimize filter parameters improve the percentage of factually correct summaries from 89.5% before filtering to 94.8% after filtering. It also decreases the percentage of summaries with minor and major factual errors from 6.3% to 4.6% and from 4.2% to 1.5%, respectively. Human evaluation also shows the informativeness score improves from 98% to 99%. The final filter parameters are summarized in Table 13 of the Appendix. Our filtering mechanism can also be applied to existing datasets for quality improvements.

4 Dataset Overview and Evaluation

Table 1 presents an overview of the final dataset, which comprises 1,369,911 training examples. Please refer to Appendix A for more statistics.

Train/validation/test splits. The dataset is divided into training, validation, and testing segments, categorized by the publication date of each article (Table 1). This approach is designed to avoid the risk of news event leakage, where a summarization model might inadvertently recall and use details of a news event from the training set to generate summaries for the testing set, rather than relying on the content of the test articles themselves.

Abstractiveness. The dataset is divided equally into "extractive" and "abstractive" bins, which we denote as CCSUM-EXT and CCSUM-ABS, respectively, based on the median MINT score (0.4665).

4.1 Evaluation

We carry out automatic and human evaluation to benchmark the proposed dataset along with other abstractive summarization datasets—CNN/DM, XSum, Newsroom, and Multi-News.

4.1.1 Automatic Evaluation

The automatic evaluation results on the test split of each dataset are presented in Table 2. To ensure the CCSUM dataset is not over-fitting on the predefined filter metrics, we add SEAHORSE (Clark et al., 2023) metrics that were not part of the filtering process—Attribution for factual consistency, Comprehensible for text quality, and MainIdea for coverage, to complement other automatic metrics. These results indicate that CCSUM summaries score highest in factual consistency, are comprehensible, and capture the main points of the articles.

Factual consistency. We use BS-Precision and Attribution (Clark et al., 2023) to evaluate factual consistency. Table 2 shows that CCSUM has the highest scores on both metrics, measured on the entire dataset. Within each data subset, Newsroom-Ext has the best factuality score, closely followed by the CCSUM-Ext subset.

Abstractiveness. Abstractiveness quantifies the lexical abstraction in summaries; high MINT (Dreyer et al., 2023) and low density (Grusky et al., 2018) indicate high abstractiveness. CCSUM is the least abstractive when measured on the entire dataset. CCSUM-ABS has comparable MINT score to that of CNN/DM, and is slightly more abstractive when measured by density.

Comprehensible. We use the Comprehensible metric defined in SEAHORSE to measure if a summary is easy to understand. Summaries that are not comprehensible often feature incomplete sentence fragments or sentences marred by non-readable characters. In our analysis, all datasets, with the exception of Newsroom, received perfect scores in comprehensibility.

Dataset	Factual Consistency		Style			Content	
	BS-Precision	Attribution	MINT	Density	Comprehensible	MainIdea	BS-Recall
CNN/DM	0.740	0.658	0.623	3.645	1.000	0.709	0.501
XSum	0.639	0.230	0.877	1.094	1.000	0.168	0.420
Newsroom	0.707	0.599	0.523	9.613	0.934	0.303	0.409
- ABS	0.570	0.233	0.895	0.944	0.971	0.100	0.358
- EXT	0.841	0.938	0.099	24.459	0.846	0.479	0.471
- MIX	0.712	0.624	0.576	3.396	0.990	0.331	0.397
Multi-News	0.626	0.082	0.656	5.019	1.000	0.110	0.566
CCSUM	0.806	0.913	0.480	5.393	1.000	0.874	0.481
- EXT	0.825	0.959	0.333	7.845	1.000	0.887	0.482
- ABS	0.786	0.865	0.631	2.889	1.000	0.861	0.480

Table 2: Benchmarking abstractive summarization datasets using automatic metrics.

Dataset	Factual Consistency			Informative	Coherence
	no error	minor	major	very good	very good
CNN/DM	86.7	7.8	5.5	63.6	63.9
XSum	50.3	35.4	14.3	45.0	87.8
Newsroom	78.9	8.4	12.7	43.9	62.8
- Newsroom-ABS	56.5	12.6	30.9	30.0	54.1
- Newsroom-EXT	92.2	5.4	2.4	59.0	66.8
- Newsroom-MIX	88.0	7.2	4.8	42.6	67.6
Multi-News	71.3	13.3	15.4	78.3	87.5
CCSUM	94.9	4.2	0.9	86.7	95.8
- CCSUM-EXT	96.9	2.4	0.6	87.6	94.9
- CCSUM-ABS	92.9	5.9	1.2	85.9	96.7

Table 3: Benchmarking abstractive summarization datasets using human annotations (%).

Coverage. Both MainIdea and and BERTScore-Recall (BS-Recall) measure the extent to which summaries cover the articles' main contents. CC-SUM has the highest coverage when measured on MainIdea, and is comparable to CNN/DM when measured on BS-Recall.

4.1.2 Human Evaluation

We also conducted a human annotation, focused on three aspects: (i) *factual consistency* where the summary has no factual errors and all the details in the summary are supported by the article, (ii) *informativeness* where the summary express the main points of the article and the summary content is important and relevant, and (iii) *coherence* where the summary has good structure and flow, is easy to follow, and facts in the summary are presented in logical order. We randomly selected 1,000 validation examples from each dataset for annotation.

The main results in Table 3 are in line with the automatic evaluation results (Section 4.1.1), showing

that CCSUM exhibits high factual consistency, informativeness, and coherence. *Factual consistency:* CCSUM has the highest factual consistency rating, with 94.9% of summaries being free of factual inconsistencies, surpassing the next best dataset, CNN/DM, by a notable 8.2% in absolute terms. *Informativeness:* CCSUM has the highest informativeness rating, with 86.7% of summaries rated as "very good". The second-best in informativeness, Multi-News, receives a "very good" rating for 78.3% of its summaries. *Coherence:* CCSUM has the best coherence rating, where 95.8% of the summaries are rated as "very good'. The second-best dataset in coherence is XSum, with 87.8% of summaries rated as "very good". The full annotation results are shown in Table 14 of the Appendix.

5 Experiments

5.1 Summarization Models

We fine-tuned FLAN-T5-Large (Chung et al., 2022) models on the abstractive and extractive subsets of the CCSUM dataset, and compared with FLAN-T5-Large trained on the CNN/DM dataset. Note that although CNN/DM was included as a pretraining dataset for FLAN-T5-Large, the generated summarizes are highly extractive, suggesting the tradeoff with other pretraining tasks. In this experiment, we fine-tune FLAN-T5-Large specifically for abstractive summarization and evaluate on the CCSUM test set. We also include Pegasus (Zhang et al., 2020), Mixtral-Instruct (Jiang et al., 2024), and GPT-3.5 (OpenAI, 2023) as baselines.

Article

Sharply lower prices for gas and cheaper used cars slowed U.S. inflation in August for a second straight month, though many other items rose in price, indicating that inflation remains a heavy burden for American households.

Consumer prices surged 8.3% in August compared with a year earlier, the government said Tuesday. Though still painfully high, that was down from an 8.5% jump in July and a four-decade high of 9.1% in June. On a monthly basis, prices rose 0.1%, after a flat reading in July.

Excluding the volatile food and energy categories, so-called core prices jumped 0.6% from July to August, higher than many economists had expected and a sign of inflation's persistence.

Despite the signs of moderating price increases, inflation remains far higher than many Americans have ever experienced and is keeping pressure on the Federal Reserve, the agency tasked with keeping prices stable. The Fed is expected to announce another big increase in its benchmark interest rate next week, which will lead to higher costs for many consumer and business loans.

Inflation has escalated families' grocery bills, rents and utility costs, among other expenses, inflicting hardships on many households and deepening gloom about the economy despite strong job growth and low unemployment.

Even if inflation peaks, economists expect it could take two years or more to fall back to something close to the Fed's annual 2% target. The cost of rental apartments and other services, such as health care, are likely to keep rising in the months ahead.

Generated Summaries

FLAN-T5-large trained on CNN/DM:
Consumer prices surged 8.3% in August compared with a year earlier. That was down from an 8.5% jump in July and a four-decade high of 9.1% in June. On a monthly basis, prices rose 0.1%, after a flat reading in July .

FLAN-T5-large trained on CCSUM-EXT:
Sharply lower prices for gas and cheaper used cars slowed U.S. inflation in August for a second straight month, though many other items rose in price, indicating that inflation remains a heavy burden for American households.

FLAN-T5-large trained on CCSUM-ABS:
U.S. consumer prices rose 8.3% in August, down from a four-decade high of 9.1% in June, but the rate of inflation remained far higher than many Americans have ever experienced.

GPT-3.5-turbo:
Lower gas prices and cheaper used cars helped slow U.S. inflation in August for a second consecutive month, but many other items saw price increases, indicating that inflation remains a burden for American households.

Figure 4: Example summaries generated by different systems.

	MainIdea	BS-Recall	Concise	MINT	ROUGE-L
CNN/DM	0.787	0.654	0.901	0.271	0.300
CCSUM-ABS	0.930	0.765	0.946	0.373	0.517
CCSUM-EXT	0.956	**0.797**	**0.994**	0.086	**0.560**
Pegasus	0.672	0.677	0.880	0.163	0.280
Mixtral-Instruct	0.992	0.751	0.985	**0.544**	0.381
GPT-3.5	**0.994**	0.753	0.983	0.458	0.415

Table 4: Automatic evaluation on generated summaries.

	Factual consistency			MINT
	no error	minor error	major error	
CNN/DM	97.2%	1.4%	1.4%	0.271
CCSUM-ABS	94.2%	4.8%	1.0%	0.343
CCSUM-EXT	98.8%	1.2%	0.0%	0.086
Pegasus	99.0%	1.0%	0.0%	0.163
Mixtral-Instruct	95.8%	3.4%	0.8%	0.544
GPT-3.5	95.6%	3.4%	1.0%	0.458

Table 5: Human evaluation on generated summaries.

5.1.1 Observations and Standard Evaluation

Observation 1. CCSUM *training leads to comprehensive summaries.* Models trained on CC-SUM can comprehensively capture the main ideas of an article while avoiding unnecessary details. Automatic evaluation results in Table 4 demonstrate that CCSUM-trained models outperform those trained on CNN/DM in capturing the main ideas, as measured in MainIdea (Clark et al., 2023) and BERTScore-Recall (Zhang et al., 2019). CC-SUM-trained models also have high Concise (Clark et al., 2023) scores. A comparative analysis in Figure 4 reveals that the summary produced by the CNN/DM-trained model misses the specific subject of the article, such as mistaking "U.S. consumer prices" for generic "consumer prices". In contrast, summaries generated by CCSUM-trained models clearly state the subject and capture the important point that "the rate of inflation remained far higher than many Americans have ever experienced." We find CCSUM-trained models better adhere to the journalistic writing style. More comparative summaries can be found in Appendix H.

Observation 2. CCSUM-ABS *yields more abstractive summaries than CNN/DM and Pegasus.* Although CCSUM-ABS has similar abstractive-

ness level with CNN/DM on the training data as shown in Table 2, models trained on CCSUM-ABS tend to generate more abstractive and concise summaries than that of CNN/DM according to the MINT (Dreyer et al., 2023) metric in Table 4. The sentence structure from CCSUM-ABS trained models is notably more complex than the simplistic language style of CNN/DM. CCSUM-ABS trained models also generate more concise summaries according to the Concise (Clark et al., 2023) metric in Table 4. Compared with Pegasus, the CCSUM-ABS models are better at capturing the main idea of an article and produce more concise summaries. We find that summaries generated by Mixtral-Instruct are notably more abstractive than those generated by other models. Mixtral-Instruct demonstrates a strong capacity in capturing the main ideas of an article while maintaining brevity. Note that details about Mixtral's pre-training or instruction-tuning data have not been disclosed.

Observation 3. CCSUM-EXT *training leads to highly-extractive summaries.* The automatic abstractiveness evaluation in Table 4 indicates that CCSUM-EXT-trained models output highly extractive summaries. CCSUM-EXT can be used as a

dataset for investigating generation-based extractive summarization models. It eliminates the need for sentence segmentation and is more robust to input noise, illustrated in Figure 13 of the Appendix.

Observation 4. Human annotations on factual consistency reveal the factuality-abstractiveness tradeoff. Table 5 summarizes the human annotation results on factual consistency. Our findings indicate that summaries generated by CCSUM-EXT are the most factual. Summaries generated by CCSUM-ABS show a similar level of major factual errors as those produced by GPT-3.5, but they exhibit a higher rate of minor factual errors. Mixtral-Instruct exhibits similar factual consistency ratings compared to GPT-3.5, where both models contain about 1% major factual error and 3.4% minor factual error. The CNN/DM-trained model has a higher factual correctness rate (92.2%) than CCSUM-ABS (94.2%) and GPT-3.5 (95.6%). Notably, there is a correlation between the level of abstractiveness (MINT) in a summary and its factual error rate. This correlation suggests that more abstractive summaries tend to have higher rates of factual errors, a finding consistent with previous research (Dreyer et al., 2023).

5.1.2 Human Preference Evaluation

We conduct a human preference evaluation to complement the results in Section 5.1.1. We ask human annotators to decide which of two summaries of a news article is better; see Appendix F for details. We first compare summaries between CCSUM-ABS and CNN/DM trained models. The findings (Table 6) show a clear preference for summaries generated by the CCSUM-ABS-trained model. This preference is shown both on the CNN/DM and the CCSUM test sets, though it is more pronounced on the CCSUM test set, hinting at a potential domain shift (Krishna et al., 2023) between these two datasets.

We also evaluate how summaries generated by models trained on CCSUM-ABS and CNN/DM fare against those produced by GPT-3.5. According to the results in Table 7, both the CNN/DM- and CCSUM-ABS-trained models are significantly less favored compared to GPT-3.5-generated summaries when the base model is FLAN-T5-Large. Increasing the size of the pre-trained model significantly narrows the performance gap between FLAN-T5 and GPT-3.5. Notably, human evaluators exhibited a slight preference for summaries generated by the FLAN-T5-XL (CCSum-Abstractive)

Test set	Preferred model	Overall	Content	Style
CCSUM	Tie	5.2%	8.8%	7.2%
	CCSUM-ABS	53.6%	39.8%	66.8%
	CNN/DM	41.2%	51.4%	26.0%
CNN/DM	Tie	3.6%	7.4%	5.2%
	CCSUM-ABS	49.0%	47.0%	46.4%
	CNN/DM	47.4%	45.6%	48.4%

Table 6: Human preference between CCSUM-ABS and CNN/DM trained FLAN-T5-Large models.

over those produced by GPT-3.5. This is particularly noteworthy considering GPT-3.5 has been fine-tuned using human preferences and might be much larger than FLAN-T5-XL. The preference for FLAN-T5-XL (CCSum-Abs) underscores the value of the proposed CCSum dataset and its journalistic-style summaries. Moreover, the results suggest that the large-scale CCSum better benefits from increases in model size, narrowing the overall preference gap from 45% (70%-25%) to -3% (47%-50%). In contrast, CNN/DM reduces the overall preference gap from 63% (79%-16%) to 51% (74%-23%). Example preference annotations between GPT-3.5 and the CCSUM-trained FLAN-T5-XL model are showcased in Appendix H.1. Given that GPT-3.5 is closed-source without transparency into its training data and model size, the proposed CCSUM dataset enables the summarization research community to train better open-source models that can outperform its closed-source counterparts. The proposed CCSUM dataset can also be used as an instruction-tuning dataset to train better instruction-following summarization models. We show in Appendix J that a controllable-summarization model can be trained from CCSUM-ABS to generate summaries at different abstractiveness levels.

6 Conclusion

We present a large-scale and high-quality dataset, CCSUM, for supervised abstractive news summarization. It contains 1.3 million pairs of articles and reference summaries derived from 35 million news articles from CommonCrawl News. In creating this dataset, we cluster CommonCrawl News articles into news events from which we generate candidate article-summary pairs and apply strict filtering and a Bayesian optimization method that eliminates 99% of the candidate summaries. The human evaluation shows the proposed dataset has higher quality—in terms of factual consistency, in-

Preferred model	Overall	Content	Style
Tie	5%	10%	9%
FLAN-T5-Large (CNN/DM)	16%	21%	10%
GPT-3.5	79%	69%	81%
Tie	3%	6%	9%
FLAN-T5-XL (CNN/DM)	23%	25%	24%
GPT-3.5	74%	69%	67%
Tie	5%	8%	14%
FLAN-T5-Large (CCSum-Abs)	25%	20%	30%
GPT-3.5	70%	72%	56%
Tie	4%	6%	11%
FLAN-T5-XL (CCSum-Abs)	50%	48%	46%
GPT-3.5	47%	46%	44%

Table 7: Human preference between CCSUM-ABS, CNN/DM trained models vs. GPT-3.5.

formativeness, and coherence—than established abstractive summarization datasets. We train summarization models on CCSUM and compare them with models trained on CNN/DM. Our human evaluation shows that summaries generated from CCSUM-trained models are favored over models trained on CNN/DM, reducing the preference gap to zero-shot summaries from GPT-3.5. We hope the proposed CCSUM dataset provides new opportunities for future research in abstractive summarization.

Limitations

In our work, factual consistency errors are defined as information in the summary that is not supported by the article content. We do not consider the case of factual hallucinations (Cao et al., 2022) where the summary can be factual according to world knowledge, but cannot be entailed by the article.

We propose a large-scale and high-quality dataset for abstractive summarization. However, with the advent of large language models, which are instruction-tuned across a variety of tasks, it is difficult to determine the amount of supervised data required to obtain high-quality abstractive summarization models. The proposed dataset offers a new platform for text generation research, inviting further research into abstractive summarization in the context of large language models.

The proposed dataset has higher quality than established abstractive summarization datasets, yet it is important to acknowledge its imperfections. Specifically, it comprises 0.9% major factual error and 4.2% minor factual error. We hope our paper raises the awareness of quality issues in exist-

ing abstractive summarization datasets, and offers inspiration for automatic approaches to improve dataset quality. Currently, our filtering mechanism is constrained by the existing automatic summary quality metrics, and the dataset can be updated once better automatic metrics are available.

In our human preference evaluation in Section 5.1.2, we observe a strong preference towards GPT-3.5-generated summaries. Details about the training procedure for GPT-3.5 have not been published, but it is widely assumed that GPT-3.5 uses RLHF (Ouyang et al., 2022) to make the model outputs more favored by human annotators, which biases the preference annotation.

Although the proposed dataset construction approach is general and can be applied to any language, the CCSUM dataset focuses on English articles from the CommonCrawl News corpus.

Ethics Statement

The human annotations in this paper are provided by qualified Mechanical Turk workers. We provided fair pay to our annotators. For preference annotation, the workers take on average 90 seconds to complete one summary comparison. We pay the workers $0.35 per question and $0.10 bonus, which leads to a pay of $14 to $18 per hour.

Like other text generation systems, generative summarization systems can suffer from hallucinations, potentially leading to misinformation. We carefully evaluate the factual consistency of the generated summaries and use it as one of the main metrics for dataset benchmarking.

The proposed dataset is derived from the CommonCrawl News corpus. Like other news datasets, the dataset may contain biases inherent in the source material. This can be due to the nature of the news sources in CommonCrawl News.

Acknowledgements

References

James C Bezdek. 2013. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

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.

Moses S Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur P. Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2023. Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2089–2105, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Peter I Frazier. 2018. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of GPT-3. *arXiv preprint arXiv:2209.12356*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Beliz Gunel, Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2020. Mind the facts: Knowledge-boosted coherent abstractive text summarization. *CoRR*, abs/2006.15435.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Tim Head, Gilles Louppe MechCoder, Iaroslav Shcherbatyi, et al. 2018. scikit-optimize/scikit-optimize: v0. 5.2. *Version v0*, 5.

Monika Henzinger. 2006. Finding near-duplicate web pages: a large-scale evaluation of algorithms. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 284–291.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Veysel Kocaman and David Talby. 2021. Spark NLP: Natural language understanding at scale. *Software Impacts*, page 100058.

Kundan Krishna, Prakhar Gupta, Sanjana Ramprasad, Byron Wallace, Jeffrey Bigham, and Zachary Lipton. 2023. USB: A unified summarization benchmark across tasks and domains. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8826–8845, Singapore. Association for Computational Linguistics.

Harold J Kushner. 1964. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. 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.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

OpenAI. 2023. OpenAI GPT-3.5 API [gpt-3.5-turbo-0613]. Available at: https://platform.openai.com/docs/models/gpt-3-5.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.

Horst Pöttker. 2003. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*, 4:501 – 511.

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

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 379–389. The Association for Computational Linguistics.

Stefan Schweter and Sajawel Ahmed. 2019. Deep-EOS: General-purpose neural networks for sentence boundary detection. In *KONVENS*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Nakatani Shuyo. 2010. Language detection library for Java.

Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. 2021. How well do you know your summarization datasets? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3436–3449, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023. MACSum: Controllable summarization with mixed attributes. *Transactions of the Association for Computational Linguistics*, 11:787–803.

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 Dataset Statistics

A.1 Word Count Statistics

The article and summary length distribution of different abstractive summarization datasets is summarized in Table 8. The article length in CCSUM is similar to that of XSUM, and the summary length in CCSUM is longer than XSUM and Newsroom, but shorter than CNN/DM.

	Training examples	Article word count percentile			Summary word count percentile		
		25%	50%	75%	25%	50%	75%
CNN/DM	286,817	419	613	876	40	51	64
XSUM	204,045	179	295	492	18	21	24
Newsroom	995,041	276	522	839	16	22	30
MultiNews	44,972	798	1,289	2,057	169	218	260
CCSUM	1,349,911	210	307	438	26	30	35

Table 8: The article and summary length statistics of abstractive summarization datasets.

A.2 Multi-Reference Statistics

The soft clustering could produce multiple summary references for each news article. On average, each article corresponds to 1.36 summaries in our final CCSUM dataset, and 235,031 articles has more than one valid summary. Therefore, CCSUM can be used to study multi-reference summarization.

A.3 Domain Statistics

The final CCSUM dataset is comprised of 4,388 unique domains. The top 15 domains and their article count in the training set are summarized in Table 9.

Domain	Article Count
www.dailymail.co.uk	63313
www.devdiscourse.com	42464
nationalpost.com	33907
www.reuters.com	33458
www.startribune.com	17688
www.channelnewsasia.com	15036
www.houstonchronicle.com	14763
www.seattletimes.com	13579
www.foxnews.com	12744
wtop.com	12721
www.washingtontimes.com	12573
www.latestly.com	12403
www.washingtonpost.com	12344
www.ctvnews.ca	10956
indianexpress.com	10354

Table 9: Top 15 domains in the training set.

A.4 Distribution of Automatic Metrics

Figure 5, 6, and 7 shows the distribution of MINT, BERTScore-Precision, and BERTScore-Recall scores in CCSUM, respectively.
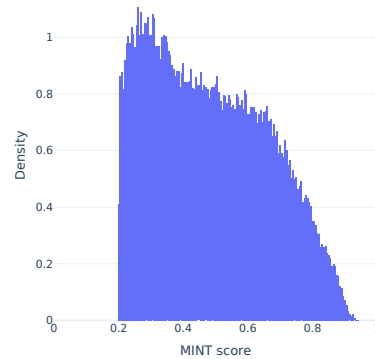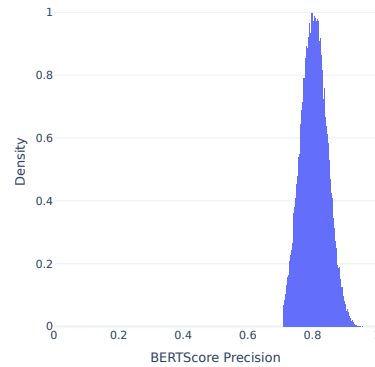

Figure 5: Distribution of the MINT scores in CCSUM.

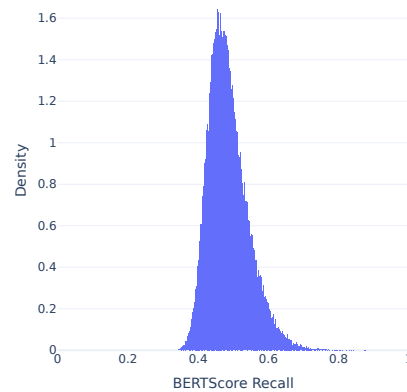Figure 6: Distribution of the BERTScore-Precision scores in CCSUM.

Figure 7: Distribution of the BERTScore-Recall scores in CCSUM.

B Example Summaries in CCSUM

Table 10 and Table 11 shows examples summaries in CCSUM-ABS and CCSUM-EXT, respectively.

Article	Summary
GADSDEN, Ala. – A fire has destroyed the home of a woman who accused U.S. Senate candidate Roy Moore of sexual misconduct. Authorities say, however, that there is no indication the blaze had anything to do with the allegations. Moore accuser Tina Johnson of Gadsden lost her home Wednesday in a fire that's under investigation by arson specialists in Etowah County. READ MORE: 2 more women accuse Republican Senate candidate Roy Moore of sexual misconduct A statement from the sheriff's office says authorities are speaking to a person of interest about the fire. The statement says investigators don't believe the fire is linked to Moore or the allegations against him. Johnson is among the women who publicly accused Moore of sexual misconduct. She told AL.com Moore groped her in his law office in 1991. Moore denied any wrongdoing, but he lost the race to Democrat Doug Jones.	The home of Tina Johnson, who accused former U.S. Senate candidate Roy Moore of groping her, was destroyed Wednesday in a fire that is being investigated as arson, though officials say they do not believe it is related to the Moore allegations.
Man charged with terrorism in stopping of Amtrak train LINCOLN, Neb. A Missouri man suspected of stopping an Amtrak train in southwest Nebraska has been charged with terrorism. Documents unsealed Wednesday in U.S. District Court in Lincoln say 26-year-old Taylor Wilson, of St. Charles, Missouri, has ties to a white supremacist group. He's charged with terrorism attacks and other violence against railroad carriers and mass transportation systems. His attorney in a related state case didn't immediately return a call Friday from The Associated Press. The eastbound train with about 175 people aboard halted early Oct. 22 in Oxford, Nebraska, about 200 miles southwest of Omaha. Court documents say a Furnas County deputy sent to the scene found Amtrak employees holding Wilson. The deputy says he found a loaded handgun in Wilson's waistband.	An armed man believed to have ties to white supremacist groups has been charged with terrorism after he stopped an Amtrak train in Nebraska.
UNITED NATIONS — The U.N. Security Council is set to hold an emergency meeting about Iran as the U.S. seeks to show support for anti-government protests in the Islamic Republic. The U.S. asked for the session, set for Friday afternoon. It's not yet clear what might come out of it. Members are divided in their views of the protests and could call a vote on whether to address the topic. U.S. Ambassador Nikki Haley has praised the demonstrators, saying "the U.N. must speak out" on their behalf. But Russia has warned against "external interference" in what it sees as an Iran's internal affairs. At least 21 people have been killed and hundreds arrested in a week of anti-government protests and unrest. Meanwhile, thousands of people took part in pro-government rallies Wednesday and Thursday. Copyright 2017 The Associated Press. All rights reserved. This material may not be published, broadcast, rewritten or redistributed. Sign up for BREAKING NEWS Emails privacy policy Thanks for subscribing!	The U.N. Security Council is planning an emergency meeting about Iran on Friday, after the U.S. asked the world body to show support for Iran's anti-government protesters.
Arsenal boss, Arsene Wenger, has insisted he is not planning to buy Pierre-Emerick Aubameyang as replacement for Alexis Sanchez. Sanchez has been linked with a move to Premier League rivals Manchester City this January, with his current deal set to expire at the end of the season. There have been speculations that Wenger will make a move for Aubameyang, who has scored 21 goals in all competitions this season, if the Chilean leaves. Wenger is, however, adamant it is not on the cards this month. "No, that is not a possibility," he told a press conference on Friday. Wenger also added that he is hopeful of keeping Theo Walcott, who has been linked with a return to Southampton. "I want him to stay. I heard about that but I want him to stay," he said.	Arsenal head coach Arsene Wenger has denied reports suggesting that the club will make a move for Pierre-Emerick Aubameyang in the current transfer window.
BANGKOK - Six Myanmar soldiers were injured in an insurgent attack in northern Rakhine state, where hundreds of thousands of Rohingya Muslims have fled to Bangladesh since the army launched a crackdown in August following militant attacks on police posts, officials said. The military said in a statement posted on the commander in chief's Facebook page that the attackers were from the Arakan Rohingya Salvation Army, the militant group blamed for the attacks on police posts. More than 20 insurgents using improvised explosives attacked the soldiers' truck, which was coming from Taungpyo township in northern Rakhine on Friday, the government said in a statement on its Facebook page. Six injured soldiers were taken to a military hospital, border guard police official Sann Oo said by phone Saturday. In the past, the military has retaliated against Rohingya following such attacks. A campaign of killings, rape and arson attacks by security forces and Buddhist-aligned mobs has sent more than 850,000 of Myanmar's 1.3 million Rohingya fleeing the country in recent years. Since Aug. 25 alone, when Myanmar's army began what it called "clearance operations" following the attacks on police posts, more than 650,000 Rohingya have fled to neighboring Bangladesh. The United Nations and the United States accuse Myanmar's military of human rights violations against Rohingya in Rakhine, including killings, rapes and the burning of homes. The U.N. has condemned the violence as ethnic cleansing.	Six Myanmar soldiers were wounded in an insurgent attack in northern Rakhine state, where government troops have been accused of "ethnic cleansing" that forced hundreds of thousands of Rohingya Muslims to flee into Bangladesh.

Table 10: Example summaries and their corresponding article in CCSUM-ABS.

Article	Summary
HELSINKI—A man in Stockholm picked up a suspected hand grenade from the ground and it detonated in his hand Sunday, killing him and injuring his companion, Swedish police said. The blast took place about 11 a.m. just outside the Varby Gard subway station in Huddinge, a residential district in greater Stockholm, said regional police spokesman Sven-Erik Olsson. "The man was seriously injured after he picked up something from the ground and this device exploded," Olsson said. The man, in his 60s, was rushed to hospital but later died while the woman, in her mid-40s, received minor wounds to her face and both legs, Olsson said. The couple had been cycling past the device when the man stopped to investigate it. Police said fragment damages on the victims and findings at the scene indicated the explosive could be a hand grenade, possibly an old one.	A man in Stockholm picked up a suspected hand grenade from the ground that detonated in his hand Sunday, accidentally killing him and injuring his female companion near a subway station.
CHICAGO A 26-year-old transgender Chicago woman serving a 10-year sentence for burglary is seeking a transfer from a male to a female prison where she says she'll be less vulnerable to abuse. A federal judge will hear testimony Friday on Deon "Strawberry" Hampton's transfer request. She says guards single her out for regular sexual abuse, taunting and beatings at the high-security men's prison she's at now in southern Illinois. Prison officials can assign such male-to-female transgender inmates to women's prisons, but it happens infrequently. Federal data from 2016 indicates there were no transgender prisoners in Illinois' two female prisons. There were 28 in the state's 24 male prisons. Testimony in Benton will include experts describing how Hampton has identified as female since she was 5.	A 26-year-old transgender woman serving a 10-year sentence in Illinois for burglary is seeking a rarely granted transfer to a female prison where she says she'll be less vulnerable to the kinds of sexual assault, taunting and beatings she's been subjected to in male prisons.
Washington Special counsel Robert Mueller is aware of an unsuccessful attempt by President Donald Trump to lobby Attorney General Jeff Sessions not to recuse himself from the Justice Department's Russia probe, The New York Times reported Thursday. The Times, citing two people with knowledge of the episode, said Trump had ordered White House counsel Don McGahn last March to stop Sessions from recusing himself from oversight of the investigation into Russian interference in the 2016 election. At the time, Sessions faced mounting calls for recusal given his support for the Trump campaign and the revelation of an omission during his confirmation hearing about contacts with Sergey Kislyak, who was the Russian ambassador to the US at the time. While widely reported on, the government did not confirm the existence of the investigation until then-FBI Director James Comey told the House Intelligence Committee later in March 2017 that there was an ongoing investigation into potential coordination between Trump's associates and Russia to influence the 2016 election. Sessions announced he would recuse himself from all matters relating to the investigation in early March. Politicians, including Republicans, praised the decision at the time, but Trump has publicly rebuked his attorney general and said he wished Sessions had not recused. Read More	President Donald Trump directed his White House counsel to tell Attorney General Jeff Sessions not to recuse himself from the Justice Department's investigation into potential ties between Russia and the Trump campaign.
Uber co-founder and former CEO Travis Kalanick plans to sell 29 percent of his stake in the ride-hailing service, a news report said Thursday. The deal is part of a transaction with investors including Softbank Group Corp. and would bring Kalanick about $1.4 billion, Bloomberg reported, citing unidentified sources. It said Kalanick previously said he never had sold Uber shares. Bloomberg didn't respond to an email seeking comment. Kalanick, who owns 10 percent of Uber, resigned as CEO last year following revelations of sexual harassment in the company, technological trickery designed to hinder regulators and a cover-up of a hacking attack that stole personal information of 57 million passengers and 600,000 drivers. Bloomberg said Kalanick offered to sell up to half his stake but reduced that due to limits in the agreement between Uber and the buyers. Uber was valued around $68.5 billion during a 2016 capital investment, but that dropped to somewhere above $48 billion in the SoftBank deal announced last week. Despite that, early investors stand to make significant gains.	Uber co-founder and former CEO Travis Kalanick will sell 29 percent of his stake in the ride-hailing service, a person briefed on the deal says.
The commander of a Colorado sheriff's deputy killed by a man with a history of mental health problems says the deputy tried to calmly help the gunman before being shot. Douglas County Sheriff Tony Spurlock spoke Friday at the funeral of 29-year-old Zackari Parrish in the Denver suburb of Highlands Ranch. Relaying what he heard on audio from body camera footage, Spurlock said Parrish pleaded with the gunman to let him help. Spurlock said he never heard anyone deal with such a situation more calmly. Parrish and three other deputies were shot after being called to the home of Matthew Riehl, an Army veteran and former Wyoming lawyer, for the second time early New Year's Eve. Parrish's funeral drew law enforcement officers from around Colorado and other states and followed a large procession on Interstate 25. People, some holding American flags, lined up to watch.	A Colorado sheriff's deputy killed by a man with a history of mental health problems tried to help the gunman before being shot, his boss said Friday at a funeral that drew law enforcement officers from around the state and elsewhere.

Table 11: Example summaries and their corresponding article in CCSUM-EXT.

C Additional Implementation Details

C.1 Data Preparation

C.1.1 Common Crawl News Data

We started the dataset construction from the Common Crawl News (CC-News)[3] dataset, which contains news articles from news sites worldwide. We use NewsPlease (Hamborg et al., 2017) to download the full Common Crawl News archive spanning from January 2018 to December 2022.

C.1.2 English Language Filter

We focus on the English subset of CC-News by filtering each article by its HTML meta-data to ensure the "language" tag is "en". However, we discovered many non-English articles are tagged as "en". To better filter articles by language, we apply an additional domain filter to identify and remove non-English domains. Among the 21,794 domains in CC-News, we randomly sample 10 articles from each domain, and assign the language as English if all articles from a domain are predicted as English with over 0.95 confidence by "langdetect" (Shuyo, 2010). This resulted in 9,999 English content domains with 58 million news articles in total.

C.1.3 Title and Content Filter

We apply a length filter to ensure each article's title is between 5 and 25 words and the main text contains at least 50 words. We also remove exact duplicates when two articles contain the same main text or when two articles have the same title and the first 200 characters of the main text are also identical. This process resulted in 35 million news articles.

C.1.4 Lead Sentence Detection and Cleaning

We use convolutional neural networks (Schweter and Ahmed, 2019) to detect sentence boundaries of each news article a_i and extract the first sentence a_i as a candidate summary for a different article $\{a_j \mid j \neq i\}$. This is implemented using Spark NLP (Kocaman and Talby, 2021) on a cluster of Amazon EC2 instances. We also use extensive regular expressions to remove datelines at the beginning of each lead sentence, such as a location, time or news providers. The primary purpose of the dateline is to indicate where and when a news story originated, and removing it does not interfere with understanding the news article.

---

[3]https://commoncrawl.org/blog/news-dataset-available

C.2 Clustering

We first encode the main text of each article into a semantic embedding using sentence transformers (Reimers and Gurevych, 2019) (all-mpnet-base-v2). We then perform soft clustering with faiss (Johnson et al., 2019) to group articles into clusters where every article's similarity to the centroid article is greater or equal to 0.9 in cosine similarity. We set the similarity threshold to 0.9 to ensure articles in the same cluster are semantically similar. Note that in soft clustering, an article can be assigned to multiple clusters, which allows us to over-generate candidate article-summary pairs that can be further filtered to obtain a large-scale and high-quality abstractive news summarization dataset.

D Filter details

D.1 CCSUM-specific Filter Heuristics

Domain difference f_{domain}. We require the candidate summary to come from a different domain than the article. Articles about a news topic from the same domain tend to be updates of the same event and could contain inconsistencies.

Summary entity count $f_{\text{entity\_count}}(s_j)$: the summary is required to have at least one entity; otherwise, the lead sentence could be a generic introductory sentence that does not capture the main points of the article.

Summary word count $f_{\text{word\_count}}$. We require the summary to contain at least 25 words.

Quotation exact match $f_{\text{quotation}}$. We use regular expression to detect quotations from the summary and ensure they are present as is in the article.

In addition to the above-mentioned filters, we also ensure the candidate summary ends in proper punctuation.

D.2 Filter Final Parameters

The full set of filters are defined in Table 12, where filter constraints with λ denote parameters to be optimized using Bayesian optimization. The final filters after optimization is summarized in Table 13. In our experiments, we apply filters in desceding order of computational complexity for computational efficiency.

D.3 Detailed Filter Funnel

Figure 8 depicts a detailed filtering funnel, including the individual heuristics and BERTScores with different base models.

Metric name	Notation f_x	Value Range	Filter Constraint A_x
Entity precision	f_{ep}	$[0, 1]$	$\{1\}$
BERTScore-Precision	$f_{\text{BS-P}}$	$[0, 1]$	$[\lambda_{\text{BS-P}}, 1]$
BERTScore-Recall	$f_{\text{BS-R}}$	$[0, 1]$	$[\lambda_{\text{BS-R}}, 1]$
Title-title similarity	$f_{\text{t-t}}$	$[0, 1]$	$[\lambda_{\text{t-t}}, 1]$
Summary-title similarity	$f_{\text{s-t}}$	$[0, 1]$	$[\lambda_{\text{s-t}}, 1]$
MINT	f_{MINT}	$[0, 1]$	$[0.2, 1]$
Simhash	f_{simhash}	$[0, 64]$	$(5, 64]$
Domain difference	f_{domain}	$\{\text{True, False}\}$	$\{\text{True}\}$
Summary entity count	$f_{\text{entity\_count}}$	$[0, \infty)$	$[1, \infty)$
Summary word count	$f_{\text{word\_count}}$	$[1, \infty)$	$[25, \infty)$
Quotation exact match $f_{\text{quotation}}$	$f_{\text{quotation}}$	$\{\text{True, False}\}$	$\{\text{True}\}$

Table 12: Overview of summary metrics and quality filters.
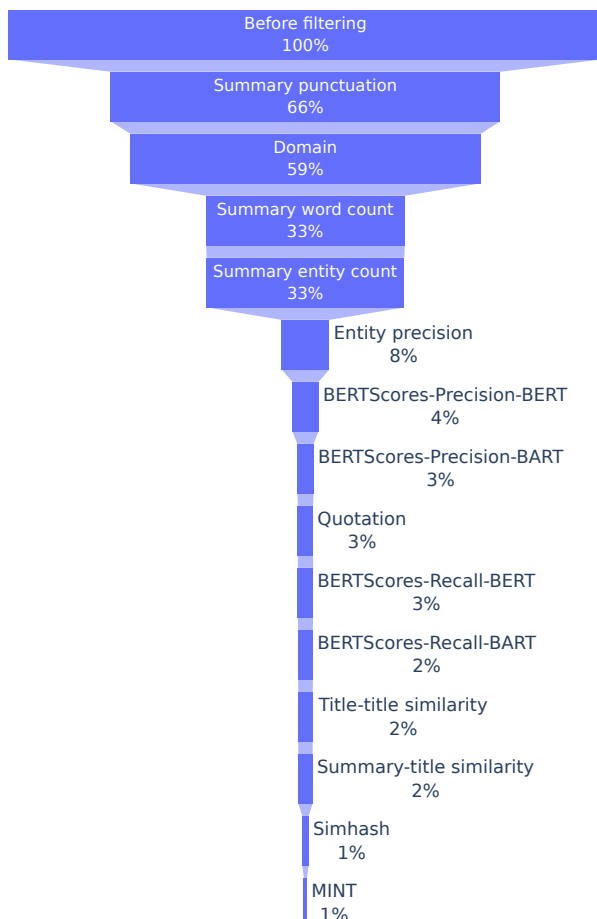
Figure 8: Detailed filter funnel.

Metric name	Constraint
Entity precision	$\{1\}$
BERTScore-Precision (bert-large-uncased)	$[0.708, 1]$
BERTScore-Precision (facebook/bart-large)	$[0.750, 1]$
BERTScore-Recall (bert-large-uncased)	$[0.344, 1]$
BERTScore-Recall (facebook/bart-large)	$[0.312, 1]$
Title-title similarity	$[0.361, 1]$
Summary-title similarity	$[0.375, 1]$
MINT	$[0.2, 1]$
Simhash	$(5, 64]$
Domain difference	$\{\text{True}\}$
Summary entity count	$[1, \infty)$
Summary word count	$[25, \infty)$
Quotation exact match	$\{\text{True}\}$

Table 13: Final summary filter parameters.

D.4 Bayesian Optimizatin Implementation Details

The optimization is carried out using Scikit-Optimize (Head et al., 2018) with the "gp_hedge" acquisition function that probabilistically chooses one of three acquisition functions at every iteration. We performed 1,000 iterations of optimization. When evaluating the filter configuration at each iteration, we return a negative constant if the constraints are violated and return $f_{\text{recall}}(\lambda, \hat{\mathcal{D}}_{\text{labelled}})$ otherwise.

E Detailed Dataset Annotation Results

Table 14 presents the detailed annotation results on the test split of each dataset. For informativeness and coherence annotation, five to one stars denote "very good", "good", "acceptable", "poor", and "very poor", respectively.

Dataset	Factual Consistency			Informativeness					Coherence				
	no error	minor	major	5-star	4-star	3-star	2-star	1-star	5-star	4-star	3-star	2-star	1-star
CNN/DM	86.7	7.8	5.5	63.6	29.7	5.7	0.7	0.3	63.9	28.9	6.5	0.5	0.2
XSum	50.3	35.4	14.3	45.0	33.2	14.5	4.0	3.3	87.8	9.4	1.5	0.4	0.9
Newsroom	78.9	8.4	12.7	43.9	23.3	19.6	7.5	5.7	62.8	20.3	11.5	3.7	1.7
Newsroom-ABS	56.5	12.6	30.9	30.0	18.6	23.7	13.8	13.8	54.1	20.4	13.8	7.5	4.2
Newsroom-EXT	92.2	5.4	2.4	59.0	24.0	12.6	3.3	1.2	66.8	18.6	13.5	0.9	0.3
Newsroom-MIX	88.0	7.2	4.8	42.6	27.3	22.5	5.4	2.1	67.6	21.9	7.2	2.7	0.6
Multi-News	71.3	13.3	15.4	78.3	17.5	3.4	0.1	0.7	87.5	9.5	2.6	0.1	0.3
CCSUM	94.9	4.2	0.9	86.7	10.6	1.9	0.8	0.0	95.8	3.2	0.9	0.1	0.0
CCSUM-EXT	96.9	2.4	0.6	87.6	9.6	2.4	0.4	0.0	94.9	3.5	1.4	0.2	0.0
CCSUM-ABS	92.9	5.9	1.2	85.9	11.6	1.4	1.2	0.0	96.7	2.9	0.4	0.0	0.0

Table 14: Benchmarking abstractive summarization datasets using human annotations (%). Five to one stars denote "very good', "good', "acceptable", "poor", and "very poor", respectively.

F Mechanical Turk Annotation Guidelines

In our Mechanical Turk annotations, we always assign three qualified annotators for each task and use their median as the final rating. The annotators are recruited from U.S., U.K., Australia, New Zealand, and Canada. For annotator qualification, we follow the procedure described in Dreyer et al. (2023). For preference annotation, the workers take on average 90 seconds to complete one summary comparison. We pay the workers $0.35 per question and $0.1 bonus, which leads to a pay of $14 to $18 per hour.

F.1 Factual Consistency

Figure 9 shows an example Mechanical Turk factual consistency annotation task, where an annotator is asked to annotate a candidate summary with respect to an article. We use sentenceBERT to highlight relevant article sentences in blue to assist the annotator. If an example is annotated as "Minor factual error" or "Major factual error," the annotator is required to explain the factual error in a few words before being able to submit the annotation.

F.2 Informativeness and Coherence

Figure 10 shows an example Mechanical Turk informativeness and coherence annotation. If an example is annotated as less than perfect rating, i.e., five stars, the annotator is required to explain why the summary is not informative or coherent.

Informative summaries express the main points of the articles; their content is important and relevant. You can give 1 to 5 stars:

- 1 star: Very poor. The summary does not express the main points of the articles. Summary content is unimportant or not relevant.

- 2 stars: Poor. The summary mostly does not express the main points of the articles. Summary content is mostly unimportant or not relevant.

- 3 stars: Acceptable. The summary expresses some main points of the articles. Some parts of the summary are important or relevant, others are not.

- 4 stars: Good. The summary mostly expresses the main points of the articles. Summary content is mostly important and relevant.

- 5 stars: Very good. The summary fully expresses the main points of the articles. Summary content is relevant and important.

Coherent summaries have good structure and flow, are easy to follow; facts are presented in logical order. You can give 1 to 5 stars:

- 1 star: Very poor. The summary has poor structure and flow or is not easy to follow. Facts are not presented in logical order.

- 2 stars: Poor. Most parts of the summary have poor structure and flow or are not easy to follow. Most facts are not presented in logical order.

- 3 stars: Acceptable. Parts of the summary have good structure and flow, are easy to follow, others are not. Some facts are presented in logical order.

- 4 stars: Good. The summary has good structure and flow, is easy to follow. Most facts are presented in logical order.

Instructions (Click to collapse)

Please evaluate how the candidate summary is with respect to the information in the article and answer two questions.

Question. Whether the summary contains <u>factual errors</u>?

Major factual error:

- **Definition:** The article and candidate summary are talking about the same news event, but the summary includes significant extra information that is not present in the article. The candidate summary can NOT be used as a valid summary for the article.
- **Example 1:** The summary might say "protest is taking place outside the Iranian embassy and France has recalled its ambassadors", but an article only includes information about the "protest" without mentioning "recalling ambassadors".
- **Example 2:** The summary might say that "A fire broke out at 2am", but the articles don't mention the time when the fire broke out, or they mention it was during the day.

Minor factual error:

- **Definition:** Most readers would not notice the error or find it less important. If printed in a newspaper, the newspaper may not need to print a correction.
- **Example 1:** The summary might say that a celebrity couple shared a video of their daughter, but the articles says that the *mom* shared the video.
- **Example 2:** The summary might say "Lady Celia Vestey was one of Prince Harry's six godmothers", but it should be *godparents*.
- **Example 3:** The summary might say "The Game Awards will take place in Los Angeles and London", but the articles say they take place "virtually from Los Angeles and London".

No factual error:

- **Definition:** The candidate summary has no factual errors. All the details in the summary are supported by the article. (Note that it is ok if the summary omits some details from the article.)
- **Example 1:** The summary might say that a celebrity couple shared a video of their daughter, and the articles says that the coupled shared the video of their daughter on *Twitter*. The summary has no factual error because it does not need to include all details from the article, e.g., "on Twitter."

Please evaluate the **candidate summary (in blue)** with respect to the article. The sentences in the articles have been highlighted with blue background color to help you find information more quickly.

Candidate Summary:

Congress president Sonia Gandhi has sought at least three more weeks to appear before the Enforcement Directorate in connection with a money laundering case pertaining to the National Herald-AJL deal.

Article:

Congress president Sonia Gandhi, who hasn't yet recovered from Covid-19, has sought more time to appear before the Enforcement Directorate.
Sources said that Gandhi has sought at least three more weeks to appear before the agency, which has summoned the Congress chief in connection with a money laundering case pertaining to the National Herald-AJL deal.
The National Herald newspaper is published by Associated Journals Limited and owned by Young Indian Pvt Limited.

Sonia Gandhi had tested positive for Covid last week.
Her son and former party chief Rahul Gandhi has also been asked to appear before the ED on June 13 at the agency's headquarters in Delhi and he is likely to appear before it.
He was earlier summoned on June 2 but had sought time as he was abroad at the time.

Advertisement
The ED had registered a fresh case under the criminal provisions of the PMLA after a trial court in Delhi took cognizance of an Income Tax Department probe against Young Indian Pvt Ltd on the basis of a private criminal complaint filed by BJP MP Subramanian Swamy in 2013.
Swamy had accused Sonia Gandhi, Rahul Gandhi and others of conspiring to cheat and misappropriate funds with Young Indian Pvt Ltd paying only Rs 50 lakh to obtain the right to recover Rs 90.25 crore that Associate Journals Ltd owed to the Congress.

Terming the ED summons "politics of revenge," Congress leader Randeep Singh Surjewala had said: "This is a politics of vendetta and revenge by the BJP to target opposition leaders as they have done to other opponents in the country.
"
"National Herald newspaper was started in 1942.
At that time the British tried to close it, today the Modi government is also doing the same thing as the Britishers did.
Now the ED is being used for this purpose..."
Meanwhile, the Congress is preparing for a major show of strength in the national capital on June 13 when Rahul Gandhi will appear before the Enforcement Directorate.
Sources said on Wednesday that all Congress MPs, of both Lok Sabha and Rajya Sabha, and senior party leaders have been asked to reach the party headquarters on Akbar Road Monday in the morning.
The party plans to take out a march to the ED office on APJ Abdul Kalam Road in a show of support for Gandhi when he appears before the agency.

Advertisement
The party has convened a virtual meeting of general secretaries, in-charges of various states and state unit presidents on Thursday evening to finalise the protest plans.
Senior party leaders said the state units should also conduct various campaigns in protest against the ED summons.

Advertisement
Read all the Latest News , Breaking News and IPL 2022 Live Updates here.

Question: Whether the summary contains <u>factual errors</u>?

○ **Major factual error**. The candidate summary contains major factual errors or extra information that is not present in the article. For example, The summary might say "protest is taking place outside the Iranian embassy and France has recalled its ambassadors", but an article only includes information about the "protest" without mentioning "recalling ambassadors".

○ **Minor factual error**. The candidate summary contains minor factual errors. Most readers would not notice the error or find it less important. For example, the summary might say that a celebrity couple shared a video of their daughter, but the articles says that the mom shared the video.

○ **No factual error**. The candidate summary has no factual errors. All the details in the summary are supported by the article.

Explain the factual error in a few words:

[   ]

[Submit]

Figure 9: Mechanical Turk Annotation Guideline for Factual Consistency Annotations

- 5 stars: Very good. The summary has very good structure and flow, is very easy to follow. All facts are presented in logical order.

F.3 Pairwise Preference

Figure 11 shows an example Mechanical Turk preference annotation where a randomized pair of summaries is presented to the annotator.

F.4 Inter-Annotator Agreement

We use the average Cohen's Kappa to quantify inter-annotator agreement. Our findings show a fair level of annotator agreement, with a score of 0.291 on the overall metric when comparing FLAN-T5-Large (CNN/DM) with GPT-3.5. However, the agreement is slight (0.138) when comparing FLAN-T5-Large trained on CCSUM with that trained on CNN/DM.

Instructions (Click to expand)

Welcome!

We need your help on evaluating an **automatically generated** summary (see below in **green**) by comparing it to the original articles. The articles are the documents that the summary tries to summarize.

Disclaimer: This task takes at least 30 seconds to complete. You do not qualify for the bonus if you spend less time.

Please answer **two questions** about the summary.

Question 1: How do you rate the **informativeness** of the summary?

Informative summaries express the main points of the articles; their content is important and relevant.

You can give 1 to 5 stars:
- **1 star**: Very poor. The summary does not express the main points of the articles. Summary content is unimportant or not relevant.
- **2 stars**: Poor. The summary mostly does not express the main points of the articles. Summary content is mostly unimportant or not relevant.
- **3 stars**: Acceptable. The summary expresses some main points of the articles. Some parts of the summary are important or relevant, others are not.
- **4 stars**: Good. The summary mostly expresses the main points of the articles. Summary content is mostly important and relevant.
- **5 stars**: Very good. The summary fully expresses the main points of the articles. Summary content is relevant and important.

Question 2: How do you rate the **coherence** of the summary?

Coherent summaries have good structure and flow, are easy to follow; facts are presented in logical order.

You can give 1 to 5 stars:
- **1 star**: Very poor. The summary has poor structure and flow or is not easy to follow. Facts are not presented in logical order.
- **2 stars**: Poor. Most parts of the summary have poor structure and flow or are not easy to follow. Most facts are not presented in logical order.
- **3 stars**: Acceptable. Parts of the summary have good structure and flow, are easy to follow, others are not. Some facts are presented in logical order.
- **4 stars**: Good. The summary has good structure and flow, is easy to follow. Most facts are presented in logical order.
- **5 stars**: Very good. The summary has very good structure and flow, is very easy to follow. All facts are presented in logical order.

We display the summary sentences as **bullet points**, to make the summary more readable.

Please evaluate the summary shown in green below.
The articles are shown for context. See instructions above.

We first list the articles:

Bruce Willis, the popular actor known for acting in the Die Hard franchise, is stepping away from his acting career, following an aphasia diagnosis - a language disorder caused by brain damage that affects a person's ability to communicate. The 67-year-old actor rose to fame with the 1980s comedy-drama series Moonlighting and has starred in more than 100 movies across his four-decade career. However, he's best known for playing a cop in the Die Hard franchise. Bruce Willis' family issued a full statement in a social media post, that read: To Bruce's amazing supporters, as a family, we wanted to share that our beloved Bruce has been experiencing some health issues and has recently been diagnosed with aphasia, which is impacting his cognitive abilities. As a result of this and with much consideration Bruce is stepping away from the career that has meant so much to him. For the uninitiated, the most common cause of aphasia is brain damage resulting from a stroke or the blockage of a blood vessel in the brain. The loss of blood leads to brain cell death or damage in areas that control language. Needless to mention, fans across the globe has been left heartbroken after the actor's family broke the news and this is how they reacted: You can check the entire post here: "Welcome to the party, pal."

Now read the **green** summary.

Summary:
Bruce Willis, the actor best known for his work on the Die Hard films, is to retire from the industry after being diagnosed with aphasia.

Please answer the two questions about the summary in the green box!

SUMMARY

How do you rate the **informativeness** of the summary?
Informative summaries express the main points of the article; their content is important and relevant.

★★★★★

How do you rate the **coherence** of the summary?
Coherent summaries have good structure and flow, are easy to follow; facts are presented in logical order.

★★★★★

If 1 to 4 stars: Describe why do you think the summary is not as informative or coherent (5-10 words):

Submit

Figure 10: Mechanical Turk Annotation Guideline for Informativeness and Coherence Annotations

Instructions (Click to collapse)

Welcome!

In this task, you will help a reader to decide which summary of a news article is better!

Suppose the user of an artificially intelligent assistant has generated two summaries for the news article below. Now read the two answers (Summary A and Summary B) and decide which one has better **content** and **style**!

You will be asked three questions. For each question, pick one of four options:

Which summary has better **content**?
This is about the *information* contained in the summary.
Pick the summary that is better at providing accurate and factual information that is relevant to the article, without adding unnecessary details or extra information not present in the article. Note that a longer summary is not necessarily better; the better summary gives you sufficient information that you feel you can trust given the news article.
[ ] Summary A   [ ] Summary B   [ ] 🥉 Tie   [ ] 👎 Both are bad

Which summary has better **style**?
This is about *how* the summary is presented.
Pick the summary that is more fluent, natural, and coherent.
[ ] Summary A   [ ] Summary B   [ ] 🥉 Tie   [ ] 👎 Both are bad

Which summary is better **overall**?
Pick the summary that you like the most!
[ ] Summary A   [ ] Summary B   [ ] 🥉 Tie   [ ] 👎 Both are bad

For each of the three questions, try to be decisive! If you have a slight preference for Summary A or Summary B, pick it! Reserve "Tie" for cases where there really is no preference for one over the other (don't use this often). Use "Both are bad" if both answers are insufficient and should not be used.

Help a reader decide which summary of the following news article is better!

A Los Angeles judge dropped four of the 11 sexual assault charges against disgraced movie producer Harvey Weinstein on Tuesday after prosecutors said they would not pursue the counts related to one of his accusers. Superior Court Judge Lisa Lench tossed two counts of forcible rape and two counts of forcible oral copulation against Weinstein, 70, dating from 2004 to 2013. He is still charged with two counts of rape and five other sexual assault counts. Weinstein is currently serving a 23-year sentence after being convicted of committing a criminal sexual act in the first degree and third-degree rape in 2020. He's pleaded not guilty to the California charges. The dismissal of the counts came three weeks after prosecutors only mentioned four women Weinstein was accused of assaulting during their opening statements, despite a fifth accuser being referred to as Jane Doe #5 in the original indictment. The Los Angeles County District Attorney's Office has not disclosed why one woman was left out of the trial. Eight others are still expected to testify against Weinstein. That includes Jennifer Siebel Newsom, the wife of California Gov. Gavin Newsom, who testified Monday and Tuesday. "He ruined my life, I put it in a box, and I was not going to not pursue my entertainment career because of what happened," Siebel Newsom said of Weistein's alleged assault in a suite at the Peninsula Hotel in Beverly Hills in 2005. "I was so violated and I don't know how that happened," she said. "I didn't see the clues and I didn't know how to escape." "As we've gotten closer to this, and it's gotten more real, my need to clarify and be more detailed" has increased, she added. "I had everything in a box, and I've been slowly sharing a little bit at a time, because this is so painful." With News Wire Services

Now read two possible summaries: **Summary A** on the left and **Summary B** on the right.

Summary A:

A Los Angeles judge on Tuesday dismissed four of the 11 sexual assault charges against Harvey Weinstein, citing prosecutors' decision not to pursue the charges against one of his accusers.

Summary B:

Four of the 11 sexual assault charges against Harvey Weinstein have been dropped by a Los Angeles judge after prosecutors said they would not pursue the counts related to one of his accusers; Weinstein is still charged with two counts of rape and five other sexual assault counts.

Which summary has better **content**?
This is about the *information* contained in the summary.
Pick the summary that is better at providing accurate and factual information that is relevant to the article, without adding unnecessary details or extra information not present in the article.
○ **Summary A**   ○ **Summary B**   ○ 🥉 Tie   ○ 👎 Both are bad

Which answer has better **style**?
This is about *how* the summary is presented.
Pick the summary that is more fluent, natural, and coherent.
○ **Summary A**   ○ **Summary B**   ○ 🥉 Tie   ○ 👎 Both are bad

Which summary is better **overall**?
Pick the summary that you like the most!
○ **Summary A**   ○ **Summary B**   ○ 🥉 Tie   ○ 👎 Both are bad

Submit

Figure 11: Mechanical Turk Annotation Guideline for Pairwise Preference Annotations

G Example Factual Consistency Errors in Abstractive Summarization Datasets

We show summary examples from each dataset that were annotated as major factual errors.

CNN/DM validation set.   Table 15 shows a major factual error in the CNN/DM validation set.

Article	A JetBlue Airways pilot who had to be restrained during a 2012 flight after shouting about Jesus and al-Qaeda sued for more than $16 million Friday, claiming the airline failed to recognize and address his mental condition. The suit filed by former captain Clayton Osbon three years after his meltdown on a flight from New York to Las Vegas, comes amid increased scrutiny of how major airlines evaluate pilots' mental health in the wake of the allegedly intentional crash of a Germanwings . Osbon, 52, claims in the lawsuit that officials at JetBlue missed several indications that something was wrong with his mental health the day of the flight, noting he missed a pre-flight meeting for the first time in years and arrived to fly disheveled and disoriented. Scroll down for video . Restrained: Clayton Osbon, 52, pictured here being taken off a New York to Las Vegas flight after his breakdown, is suing JetBlue for failing to recognize he was mentally ill . Osbon moved slowly during pre-flight checks for the flight to Las Vegas from New York, and later relieved himself of duty after being told by the first officer he had missed several calls from air traffic control, the lawsuit adds. After a co-pilot locked Osbon out of the cockpit, he 'ran down the aisles screaming and ranting concerning imagined terrorism and the need for all on board on embrace religion.' The lawsuit alleges that JetBlue allowed the flight to continue for three hours and 'unnecessarily endangered the lives of Captain Osbon, the crew and the 135 passengers.' There were no injuries on the flight, which eventually made an emergency landing in Amarillo, Texas. The meltdown resulted from a complex partial brain seizure that 'severely impaired his ability to perform basic activities, caused him to hallucinate, and caused extreme feelings of paranoia and religious fervor,' according to the suit. The lawsuit also alleges the airline routinely ignored crew members' impairment, whether by alcohol, drugs or mental and emotional distress, in order to protect their careers. Hospitalized: Osbon, right, is led out by FBI agents after being released from Northwest Texas hospital in Amarillo following his mid-air breakdown . Veteran captain: Osbon, left, was an experienced pilot but a 'complex partial brain seizure' impaired his brain functions and caused him to hallucinate, according to the lawsuit . Nearly three dozen passengers sued JetBlue in 2012 for being 'grossly negligent' by allowing Osbon to fly with unchecked mental issues. JetBlue eventually settled the lawsuit in April of the following year for an undisclosed sum. After the flight was grounded in Texas, Osbon was indicted on charges of interfering with a flight crew and found not guilty by reason of insanity. Among other conditions of his sentence, Osbon was ordered not to board a commercial or private plane without permission from a probation officer. Osbon remains suspended on medical leave, no longer collecting what once was a more than $300,000 annual salary. The lawsuit states he has not suffered a seizure since beginning a regimen of prescription medication.
Summary	Clayton Osbon, 52, claims in a lawsuit the airline failed to recognize signs he was in mental distress before a flight in which he broke down shouting . The lawsuit claims Osbon missed a pre-flight meeting and arrived disheveled . The meltdown was caused by a seizure resulting from a lingering traumatic brain injury that occurred in Osbon's childhood, the lawsuit states . After relieving himself of duty during the flight, Osbon was locked out of the cockpit . He shouted at passengers about Jesus and al-Qaeda before being tackled and restrained .
Error	The article do not state the seizure was due to a brain injury during his childhood.

Table 15: Factual consistency error in the CNN/DM validation set.

Newsroom validation set.   Table 16 shows a major factual consistency error in the Newsroom validation set, where the article doesn't the investment is low by European standards but is included in the summary.

Article	To infinity and beyond! BT Group channeled its inner Buzz Lightyear with a headline-grabbing set of investment pledges Thursday. Its standard "superfast" fiber-broadband package, called BT Infinity, will be extended and in some places superseded by "ultrafast" services. Together with investment in improving the 4G network of EE, the mobile business it has just bought, this will cost £6 billion ($8.7 billion) over three years.
Summary	For all of BT's eye-catching pledges, investment looks low by European standards.
Error	Article doesn't mention low investment by European standards.

Table 16: Factual consistency error in the Newsroom validation set.

XSum validation set. Table 17 shows a major factual consistency error in the XSum validation set, where the article doesn't state that profits jumped 21% as in the summary.

Article	It reported underlying pre-tax profits of £381m in the year to 2 May, with revenues up 6% to £9.9bn. The firm was formed last year by the merger of Carphone Warehouse and Dixons Retail. "This has been a terrific first year for Dixons Carphone," said the firm's chief executive Sebastian James. "We have seen excellent increases in both sales and profitability and we have made very encouraging progress with the tricky job of integrating these two great companies," he continued. In the UK and Ireland, where it trades under the Carphone Warehouse, Currys and PC World names, sales rose by 8UK sales were helped by the failure of rival mobile retailer Phones4U, which collapsed last September. Keith Bowman, equity analyst at Hargreaves Lansdown Stockbrokers: "Profitability at the group's core UK and Irish operations has enjoyed a solid increase, aided by gains in market share, whilst merger cost savings remain on track." In Greece, where Dixons Carphone trades under the Kotsovolos name, the firm said it recorded an increase in like-for-like revenues, with strong demand for large screen TVs. "We have a fantastic team in Greece," Mr James told the BBC. "They've been very entrepreneurial in thinking about every possible outcome, and how we would react to it." However, trading in southern Europe had struggled with like-for-like sales down 5%. The firm said its business in Spain continued to operate in a "tough marketplace". Earlier this month, Dixons Carphone announced a deal with US telecoms firm Sprint, which could lead to a joint venture opening up to 500 stores in the US.
Summary	Dixons Carphone has reported a 21% jump in profits in its first annual results since the merger that created the mobile phone and electrical goods firm.
Error	Article doesn't state that profits jumped 21%.

Table 17: Factual consistency error in the XSum validation set.

CCSUM validation set. Table 18 shows a major factual consistency error in the CCSUM validation set. The article states four new cases of monkeypox have been identified whereas the summary states only two new cases. This is a challenging example that failed to be detected by our filters. The article mentions "two people that had previously been suspected of having monkeypox have tested negative, but four new suspected cases were reported today." One might argue the four suspected cases is an increase of two cases from the previous two suspected cases, which makes it difficult to detect the factual consistency error.

Article	Toronto Public Health has confirmed the city's first case of monkeypox.The city has been investigating several suspected and probable cases, and today announced that one has been lab confirmed.Officials say the person is in stable condition and recovering in hospital.The city also says two people that had previously been suspected of having monkeypox have tested negative, but four new suspected cases were also reported today.That means there are four people with suspected cases and one person with a probable case, and public health officials say they are all recovering at home.Monkeypox is a rare disease that comes from the same family of viruses that causes smallpox, which the World Health Organization declared eradicated around the globe in 1980, but monkeypox generally does not spread easily between people and is transmitted through prolonged close contact.This report by The Canadian Press was first published May 26, 2022.
Summary	Toronto Public Health says there are two new suspected cases of monkeypox in the city, as well as one probable case of the virus that's under investigation.
Error	Two new suspected cases is inconsistent with article which says four

Table 18: Factual consistency error in the CCSUM validation set.

H Generated Summary Examples

Table 19 shows a set of comparative example summaries generated by different systems. We find the CCSUM-ABS-trained models produce more comprehensive summaries and properly states the main story in logical order. In comparison, the CNN/DM model's summary omits the name of the suspect. When compared with GPT-3.5 outputs, we notice the CCSUM-trained model produces very similar summaries with the first half of GPT-3.5 summary, except that GPT-3.5 includes more information in the second sentence "Weinstein is still charged with two counts of rape and five other sexual assault counts."

Article	A Los Angeles judge dropped four of the 11 sexual assault charges against disgraced movie producer Harvey Weinstein on Tuesday after prosecutors said they would not pursue the counts related to one of his accusers. Superior Court Judge Lisa Lench tossed two counts of forcible rape and two counts of forcible oral copulation against Weinstein, 70, dating from 2004 to 2013. He is still charged with two counts of rape and five other sexual assault counts. Weinstein is currently serving a 23-year sentence after being convicted of committing a criminal sexual act in the first degree and third-degree rape in 2020. He's pleaded not guilty to the California charges. The dismissal of the counts came three weeks after prosecutors only mentioned four women Weinstein was accused of assaulting during their opening statements, despite a fifth accuser being referred to as Jane Doe #5 in the original indictment. The Los Angeles County District Attorney's Office has not disclosed why one woman was left out of the trial. Eight others are still expected to testify against Weinstein. That includes Jennifer Siebel Newsom, the wife of California Gov. Gavin Newsom, who testified Monday and Tuesday. "He ruined my life, I put it in a box, and I was not going to not pursue my entertainment career because of what happened," Siebel Newsom said of Weistein's alleged assault in a suite at the Peninsula Hotel in Beverly Hills in 2005. "I was so violated and I don't know how that happened," she said. "I didn't see the clues and I didn't know how to escape." "As we've gotten closer to this, and it's gotten more real, my need to clarify and be more detailed" has increased, she added. "I had everything in a box, and I've been slowly sharing a little bit at a time, because this is so painful." With News Wire Services
FLAN-T5-Large (CCSUM-ABS)	A Los Angeles judge on Tuesday dismissed four of the 11 sexual assault charges against Harvey Weinstein, citing prosecutors' decision not to pursue the charges against one of his accusers.
FLAN-T5-Large (CNN/DM)	Judge Lisa Lench tossed two counts of forcible rape and two counts of forcible oral copulation. He is still charged with two counts of rape and five other sexual assault counts. The dismissal of the counts came three weeks after prosecutors only mentioned four women.
GPT-3.5	Four of the 11 sexual assault charges against Harvey Weinstein have been dropped by a Los Angeles judge after prosecutors said they would not pursue the counts related to one of his accusers; Weinstein is still charged with two counts of rape and five other sexual assault counts.

Table 19: Comparative summaries generated by different systems (I).

Table 20 shows another set of comparative summary examples. The summary from CNN/DM states "Fina" without any explanation, whereas the CCSUM-ABS trained summaries are easier to understand by using "Swimming's world governing body." We also notice that GPT-3.5's summary is more detailed.

Article	Fina also intends to establish a new 'open' category for swimmers whose gender identity differs from their birth sex Swimming's world governing body, Fina, has voted to ban all transgender athletes that have gone through any part of male puberty from taking part in elite women's races. Fina's new policy, which passed with 71Furthermore, Fina also aims to establish an 'open' category at its competitions for swimmers whose gender identities differ from their birth sex. The new policy will require that transgender competitors must have completed their transition by the time they are 12 years old in order to take part in women's competitions. Before casting their votes, Fina members heard a report from a transgender task force containing leading figures in the worlds of sport, medicine, and law. Speaking after news of the vote broke, Fina's executive director Brent Nowicki said that the governing body had "emphasized competitive fairness" in its approach to drafting the policy, which was comprehensive, science-based and inclusive. Elsewhere, Fina president Husain Al-Musallam claimed that the global authority was trying to "protect the rights of our athletes to compete" in addition to "protecting competitive fairness". "Fina will always welcome every athlete," Al-Musallam insisted. "The creation of an open category will mean that everybody has the opportunity to compete at an elite level." "This has not been done before, so Fina will need to lead the way. I want all athletes to feel included in being able to develop ideas during this process," he added. Sharron Davies, an ex-Great Britain swimmer who has continually argued against transgender participation in women's swimming, tweeted that she was "proud" of Fina and her sport for "doing the science, asking the athletes/coaches and standing up for fair sport for females". "Swimming will always welcome everyone no matter how you identify but fairness is the cornerstone of sport," she stressed. But LGBT advocacy group 'Athlete Ally' branded the new policy "discriminatory, harmful, unscientific and not in line with the 2021 IOC principles". "If we truly want to protect women's sports, we must include all women," said a tweet from the group, which has backed former UPenn college swimmer Lia Thomas in the past. Fina's new policy comes at a time when transgender participation in women's competitions is one of the most debated topics in sport and has only intensified with Thomas' success. In March, Thomas, who swam on the UPenn men's team before undergoing hormone replacement treatment in 2019, won the highest national college title in the US when taking part in the women's 500-yard freestyle, and she also broke several records at her former Ivy League college. In cycling, there has also been a furor over British cyclist Emily Bridges being allowed to take part in elite women's events or not. Moves from Fina's counterpart in the sport the UCI this week mean that she will have to wait until 2023 for this given that the body has doubled the period of time before a rider who has transitioned from male to female can take part in women's events to two years.
FLAN-T5-Large (CCSUM-ABS)	Swimming's world governing body has voted to ban transgender athletes who have gone through any part of male puberty from competing in elite women's races.
FLAN-T5-Large (CNN/DM)	Fina has voted to ban transgender athletes from taking part in women's races . The decision was made at an extraordinary general congress at the World Championships . Fina also intends to establish an 'open' category for swimmers whose gender identity differs from their birth sex .
GPT-3.5	Swimming's governing body, Fina, has voted to ban transgender athletes who have gone through any part of male puberty from competing in elite women's races, but also aims to establish an 'open' category for swimmers whose gender identity differs from their birth sex. The new policy requires transgender competitors to have completed their transition by the age of 12 in order to participate in women's competitions.

Table 20: Comparative summaries generated by different systems (II).

H.1 Human Preference between CCSUM-trained FLAN-T5-XL and GPT-3.5

H.1.1 Example Summaries where CCSUM-trained Models are Preferred over GPT-3.5

Table 21 and Table 22 showcase instances where the FLAN-T5-XL model model trained on CCSUM generates abstractive summaries that are more favored than those by GPT-3.5 across all three metrics—content, style, and overall—according to Mechanical Turk annotations.

In Table 21, both summaries are effective, but the CCSUM-based summary is slightly better. It is more specific in identifying the relationship between the perpetrator and the victim—"father" and "daughter", which is crucial in understanding the context of "honour killing." The CCSUM-based summary is also more specific in describing the relationship with "a Dalit boy" in contrast with "a Dalit guy."

In Table 22, the CCSUM-based summary provides specific and clear reasons for Bayern Munich's decision to sign Sadio Mane, making it more informative and accurate compared to the vague and less informative reason of being "a rare breed" in the GPT-3.5 summary.

Article	A man killed his daughter for loving a Dalit guy and surrendered to police in Karnataka's Mysuru district on Wednesday, authorities said, in what could be an honour killing.The tragic occurrence occurred inside the Periyapattana police station's jurisdiction. Suresh, the accused father, arrived in the police station on Tuesday and admitted to the crime, according to authorities.Suresh, a native of Kaggundi village in Mysuru district's Periyapatna taluk, murdered his 17-year-old daughter Shalini in the early hours of Monday. According to the police, the accused father strangled his daughter to death. Shalini was a second-year PUC student from the Vokkaliga community, which is considered an elite caste in Karnataka. She was in love with a Dalit kid from the nearby village of Mellahalli. According to investigators, they had been in love for the past three years. When the parents learned of it, they filed a complaint against the lad because the girl was a minor. At the police station, the girl made a statement against her parents.She told the cops she was in love with the boy and that she didn't want to go with her parents. She was taken to the Observation Home by the police. The daughter had contacted her parents and requested that they take her home some time ago. She told her parents when she got home that she still loves the lad and would only marry him.According to police, the father became enraged and strangled her to death on Monday. Later, he put his daughter's body in one of the Dalit boy's village's farmlands. The case is being looked at further.
FLAN-T5-XL (CCSUM-ABS)	A father killed his daughter for loving a Dalit boy and surrendered to police in Karnataka's Mysuru district on Wednesday in what could be an honour killing.
GPT-3.5	A man in Karnataka, India, killed his daughter for being in love with a Dalit guy and surrendered to the police, potentially constituting an honor killing.

Table 21: Example summary where CCSUM-ABS-trained models are preferred over GPT-3.5. (I).

Article	Bayern Munich chief, Oliver Kahn, has suggested why the club signed Sadio Mane from Liverpool. Bayern Munich officially confirmed the signing of Mane from Liverpool on Wednesday. Mane joined the Bundesliga champions on a three-year deal from Premier League giants Liverpool. In a tweet via his Twitter handle on Wednesday, Kahn said Mane's great successes at the international level made Bayern sign him. He tweeted: "We are happy that we could win Sadio Mané for the @FCBayern. "Thanks to his outstanding performances and his great successes at the top international level, he has been a player for many years that is very rare in the world." Mane recently won the Africa Cup of Nations, AFCON, with the Senegal national team and qualified them for this year's FIFA World Cup in Qatar."
FLAN-T5-XL (CCSUM-ABS)	Bayern Munich chief Oliver Kahn has said that Sadio Mane's outstanding performances and international success were the reasons behind the club signing him from Liverpool.
GPT-3.5	Bayern Munich president Oliver Kahn has revealed the reason the club decided to sign Liverpool forward Sadio Mane on Wednesday, stating that the player is a rare breed.

Table 22: Example summary where CCSUM-ABS-trained models are preferred over GPT-3.5. (II).

H.1.2 Example Summaries where GPT-3.5 is Preferred over CCSUM-trained Models

Table 23 and 24 showcase instances where GPT-3.5 summaries are more favored than CCSUM-trained FLAN-T5-XL models across all three metrics—content, style, and overall—according to Mechanical Turk annotations.

In Table 23, the summary from FLAN-T5-XL (CCSUM) is nearly the same with the first half of GPT-3.5 summary, but the GPT-3.5 summary is a compound sentence where the second clause contains additional information on the context of the visit, i.e., "as the West and wider Middle East seek to repair relations with Saudi Arabia amid financial strains caused by..."

In Table 24, the GPT-3.5 summary is better because it mentions the name of the subject "Jason Alexander" rather than describing it as "the man who...". The GPT-3.5 summary is also more informative that mentions the subject "was briefly married to Britney Spears in 2004," which was not included in the summary from FLAN-T5-XL (CCSUM).

Article	By Tamara Qiblawi, CNN Saudi Crown Prince Mohammed bin Salman has landed in the Turkish capital, Ankara, for a first visit to the country since the 2018 murder of journalist Jamal Khashoggi at the kingdom's Istanbul consulate. The visit began with a welcome ceremony, and will be followed by a one-on-one meeting between Turkish President Recep Tayyip Erdogan and the crown prince. The trip comes as the West and wider Middle East seek to repair relations with the oil-rich kingdom, in a bid to alleviate financial strains prompted by the pandemic and sky-high energy prices sparked by Russia's invasion of Ukraine. In April, Erdogan met the crown prince in the Saudi city of Jeddah, a visit that ended the years-long diplomatic standoff between the two countries. The crown prince — known as MBS — is the de facto ruler of Saudi Arabia. A wave of crackdowns on Saudi dissidents, which culminated in Khashoggi's murder by a 15-man hit squad, strained relations between Riyadh and multiple Western states and Turkey. However, then-US President Donald Trump stood by MBS — who was a lynchpin of Trump's Middle East policy — even as the CIA said they believed the crown prince approved the murder at the consulate. According to Turkish official sources, a Saudi forensic expert equipped with a bone saw dismembered Khashoggi's body after he was killed in the Istanbul consulate in October 2018. MBS denied involvement in the murder, which captured international headlines with lurid details of the killing. Erdogan spoke out strongly against the murder, launching a probe into the killing, and diplomatic relations between the two countries broke down. But a plummeting Turkish lira and inflation of more than 70% appear to have forced the Turkish President to change tack. For months, Ankara has made overtures to Riyadh — as well as Saudi's closest Gulf ally, Abu Dhabi — to mend relations, including ending Khashoggi's murder trial in April and transferring the case to Riyadh. Rights groups have condemned the move, arguing it would kill off the case. Khashoggi's fiancée, Hatice Cengiz, an outspoken advocate for justice for the slain Saudi critic, said her legal team would appeal the decision. "The fact that he came to our country does not change the fact that he is responsible for murder," Cengiz tweeted on Wednesday, referring to MBS' visit. During his US presidential campaign, Joe Biden vowed to turn MBS into a "pariah" over his rights record. As President, Biden has declined to communicate directly with the powerful prince, opting instead to speak to his official counterpart — the ailing King Salman. But soaring fuel prices in recent weeks sparked a U-turn in Biden's Saudi policy. Tense diplomatic relations drastically reduced US leverage in pushing Riyadh to pump more oil, until Biden ramped up diplomatic efforts and scheduled a first official visit to the kingdom, which is expected to happen in July. He is expected to meet MBS during the visit. The-CNN-Wire ™ & © 2022 Cable News Network, Inc., a WarnerMedia Company. All rights reserved. CNN's Isil Sariyuce contributed to this report.
FLAN-T5-XL (CCSUM-ABS)	Saudi Crown Prince Mohammed bin Salman arrived in Turkey on Wednesday for his first official visit since the murder of journalist Jamal Khashoggi in the Saudi consulate in Istanbul in October 2018.
GPT-3.5	Saudi Crown Prince Mohammed bin Salman has arrived in Turkey for a visit, marking his first trip to the country since the 2018 murder of journalist Jamal Khashoggi; the visit comes as the West and wider Middle East seek to repair relations with Saudi Arabia amid financial strains caused by the pandemic and high energy prices.

Table 23: Example summary where GPT-3.5 is preferred over CCSUM-ABS-trained FLAN-T5-XL. (I).

Article	The felony charge came from multiple incidents over time, an attorney said. The pair were married for less than three days in 2004. Example video title will go here for this video VENTURA, Calif. — A man once briefly married to Britney Spears who showed up uninvited at the pop star's wedding to longtime boyfriend Sam Asghari was charged Monday with felony stalking. Jason Alexander, 40, pleaded not guilty in Ventura County court to the charge, along with misdemeanor counts of trespassing, vandalism and battery. The three misdemeanor charges stemmed directly from Alexander's attempt to get into the wedding, which he livestreamed on Instagram. The stalking charge involved repeated incidents over time, Spears attorney Mathew Rosengart said. "He's going to — I hope, and I pledged my support — be very vigorously prosecuted," Rosengart said outside the hearing. "It's a very serious matter. This was more than just a quote 'wedding crash.' This was an intrusion." A judge issued a restraining order requiring Alexander, who appeared in court via video conference from jail, to stay at least 100 yards from Spears for three years. The judge set his bail at $100,000 and scheduled a Wednesday hearing on the issue. An email seeking comment from his public defender Matthew Terry was not immediately returned. Alexander was Spears'first husband. The two were married for less than three days in 2004 before an annulment. Spears, 40, and Asghari, 28, were married Thursday at her home in Thousand Oaks, California, in front of several dozen guests including Selena Gomez, Drew Barrymore, Paris Hilton and Madonna. Alexander, who was a childhood friend of Spears, was streaming live on Instagram when he approached the house. He appeared in a mostly empty but decorated room, telling security Spears had invited him. Deputies responding to a trespassing call detained him, and he was arrested when they learned he had a warrant for his arrest in another county, the Ventura County Sheriff's Office said. Spears was previously married from 2004 to 2007 to Kevin Federline, with whom she shares two sons, ages 14 and 15. She met Asghari in 2016 on the set of the video for her song "Slumber Party." Their wedding came nine months after Spears and Asghari were engaged, and nearly seven months after Spears won her freedom from a court conservatorship that controlled her life and affairs for more than 12 years. Spears posted about the nuptials on her Instagram and Twitter accounts the following day.
FLAN-T5-XL (CCSUM-ABS)	The man who was arrested after he showed up uninvited to Britney Spears' wedding to Sam Asghari pleaded not guilty Monday to felony stalking and multiple misdemeanors.
GPT-3.5	Jason Alexander, a man who was briefly married to Britney Spears in 2004, has been charged with felony stalking after showing up uninvited at her recent wedding to Sam Asghari; he has also been charged with misdemeanor counts of trespassing, vandalism, and battery.

Table 24: Example summary where GPT-3.5 is preferred over CCSUM-ABS-trained FLAN-T5-XL. (II).

I Additional Experiment Details

I.1 Fine-tuning FLAN-T5

We use deepspeed (Rasley et al., 2020) to fine-tune FLAN-T5 models on eight Nvidia A100 GPUs. We use learning rate 1e-5 with the AdamW (Loshchilov and Hutter, 2017) optimizer. All FLAN-T5 models are fine-tuned for 2 epochs.

I.2 GPT-3.5 Prompts

We have tried different prompts to generate abstractive summaries using GPT-3.5, but many prompts lead to very long summaries that contains about one hundred words. The final prompt used in the experiment is "Summarize the following news article into one brief sentence: {article}."

I.3 Length Statistics of Generated Summaries

Table 25 summaries the output word count statistics of different summarization systems. We find that CCSUM-trained models generate longer summarizes than the pretrained FLAN-T5-Large, and shorter summarizes than that of CNN/DM or GPT-3.5.

	Percentile			
	25%	50%	75%	Std
FLAN-T5-Large	20.0	26.0	32.0	9.1
CNN/DM	35.0	41.2	47.0	10.5
GPT-3.5	32.0	41.6	49.0	13.7
CCSUM-ABS	26.0	29.0	32.0	5.1
CCSUM-EXT	27.0	31.7	36.0	7.5

Table 25: Word count statistics of generated summaries.

J Category-instruct summarization

We train a category-based instruction-following model (Ribeiro et al., 2023) to summarize a news article into different abstractiveness levels, e.g., "Summarize with abstractiveness level 8," where the requested abstractiveness level corresponds to the MINT × 10. Figure 12 depicts the relationship between the requested abstractiveness score and the actual abstractiveness score. We find that the model generates more abstractive summaries as we increase the requested abstractiveness score, but the received abstractiveness score is discounted by 25%, suggesting more research is needed to obtain more precise instruction-following summarization models. The proposed CCSUM dataset can be used
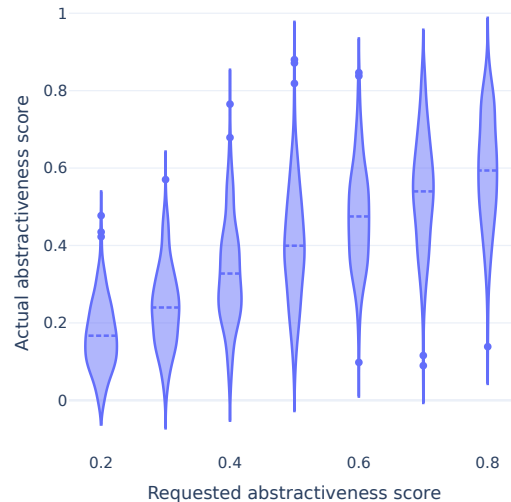

Figure 12: Requested abstractiveness score in user prompt vs. actual abstractiveness score in the summary.

in more controllable summarization tasks (Zhang et al., 2023).

K The De-noising Effect of CCSUM-EXT-trained models

Figure 13 shows that FLAN-T5-Large trained on CCSUM-EXT can effectively remove noise, in red, from the article when generating a highly extractive summary, which resolves the downsides of imperfect sentence segmentation.

Article without noise

Germany is in talks with Canada over options to export liquefied natural gas to Europe via a terminal on Canada's east coast, a German government official told Reuters on Tuesday.German Chancellor Olaf Scholz discussed the issue with Canadian Prime Minister Justin Trudeau on the sidelines of the summit held by The Group of Seven economic powers leaders this week, two other sources told Reuters. Berlin has been rushing to phase out Russian energy imports after Moscow's invasion of Ukraine and is looking for alternative supply routes and sources of energy. It has backed the construction of two LNG terminals and has rented four floating storage and regasification units as a stop-gap measure.

In May, Canadian Natural Resources Minister Jonathan Wilkinson said the Canadian government was in discussions with the companies behind two proposed east coast LNG export facilities to see how it could speed up the projects and help boost supply to Europe. Canada, the world's sixth-largest natural gas producer, does not have any east coast LNG facilities and only one under construction on its west coast. "In terms of CO_2 footprint and proximity to European markets Canadian east coast projects are very well placed," one of the sources said. The second source said that Canadians were eager to develop new fields given high shale gas prices and recognised Germany's reputation as being environmentally friendly.

Germany consumes around 100 billion cubic meters of natural gas annually with around 55% of that coming from Russia and smaller volumes piped from Netherlands and Norway. Spanish oil major Repsol's regasification plant on Canada's east coast, through which it imports into the United States, could potentially have a liquefaction capacity installed to export directly into Northwest Europe, an industry source said. The office of Canadian Prime Minister Justin Trudeau was not immediately available for comment. Repsol told Reuters it is continuously exploring options to maximize the value of the terminal, with a particular focus on new lower-carbon opportunities to help meet market demand. "The company will look at any/all business that enhances or creates value at Saint John LNG, including the potential to add liquefaction capabilities to the existing facility," it added.

Summary:

Germany is in talks with Canada over options to export liquefied natural gas to Europe via a terminal on Canada's east coast, a German government official told Reuters on Tuesday.

Article with noise

Germany is in talks with Canada this is breaking news, come back to read more updates over options to export liquefied natural gas to Europe via a terminal on Canada's east coast, a German government official told Reuters on Tuesday.German Chancellor Olaf Scholz discussed the issue with Canadian Prime Minister Justin Trudeau on the sidelines of the summit held by The Group of Seven economic powers leaders this week, two other sources told Reuters. Berlin has been rushing to phase out Russian energy imports after Moscow's invasion of Ukraine and is looking for alternative supply routes and sources of energy. It has backed the construction of two LNG terminals and has rented four floating storage and regasification units as a stop-gap measure.

In May, Canadian Natural Resources Minister Jonathan Wilkinson said the Canadian government was in discussions with the companies behind two proposed east coast LNG export facilities to see how it could speed up the projects and help boost supply to Europe. Canada, the world's sixth-largest natural gas producer, does not have any east coast LNG facilities and only one under construction on its west coast. "In terms of CO_2 footprint and proximity to European markets Canadian east coast projects are very well placed," one of the sources said. The second source said that Canadians were eager to develop new fields given high shale gas prices and recognised Germany's reputation as being environmentally friendly.

Germany consumes around 100 billion cubic meters of natural gas annually with around 55% of that coming from Russia and smaller volumes piped from Netherlands and Norway. Spanish oil major Repsol's regasification plant on Canada's east coast, through which it imports into the United States, could potentially have a liquefaction capacity installed to export directly into Northwest Europe, an industry source said. The office of Canadian Prime Minister Justin Trudeau was not immediately available for comment. Repsol told Reuters it is continuously exploring options to maximize the value of the terminal, with a particular focus on new lower-carbon opportunities to help meet market demand. "The company will look at any/all business that enhances or creates value at Saint John LNG, including the potential to add liquefaction capabilities to the existing facility," it added.

Extractive summary removes noise:

Germany is in talks with Canada over options to export liquefied natural gas to Europe via a terminal on Canada's east coast, a German government official told Reuters on Tuesday.

Figure 13: FLAN-T5-Large trained on CCSUM-EXT can effectively remove noise, in red, from the article when generating a highly extractive summary, which resolves the downsides of imperfect sentence segmentation.

Sentence	CCSum-Abs	CCSum-Ext	GPT-3.5	CNN/DM
1	68.8%	68.6%	73.0%	31.8%
2	18.2%	24.0%	16.2%	16.4%
3	5.2%	3.6%	5.4%	10.2%
4	2.2%	1.2%	2.4%	7.0%
5	2.2%	0.6%	1.0%	8.0%
6	1.2%	1.0%	0.4%	6.8%
7	1.2%	0.2%	0.8%	6.6%
8	0.6%	0.2%	0.2%	4.6%
9	0.4%	0.6%	0.6%	4.2%
10	0.0%	0.0%	0.0%	4.4%

Table 26: The distribution of the most relevant sentence index in the article w.r.t. the summary.

L Similarity between Reference Summary and Article Sentences

In Table 26, we have measured the distribution of the most relevant sentence (limited to the first ten sentences for simplicity) in relation to the summary. This measurement is based on cosine similarity using Sentence-BERT. Additionally, in Table 27, we've included the average cosine similarity between the reference summary and each of the first ten sentences of the article. To establish a baseline, we also calculated these statistics using summaries generated by GPT-3.5.

Our findings indicate that in the CCSum-Abs dataset, the reference summaries most closely resemble the first sentence of the article in 68.8% of cases. This is marginally lower than the 73% observed with GPT-3.5 summaries where the first sentence is most often the most relevant. Furthermore, our data shows a monotonous decrease in cosine similarity between the reference summary and the sentences in the article. These observations suggest that the CCSum dataset exhibits similar characteristics to GPT-3.5 in terms of the positional relevance of sentences in the article. We also report the statistics for the CNN/DM dataset by encoding the three-sentence summary into a single vector. The findings suggest that CCSum has a stronger lead-bias than CNN/DM. The positional relevance in CNN/DM is more distributed across various sentences, and there is a comparable monotonic decrease in cosine similarity.

Sentence	CCSum-Abs	CCSum-Ext	GPT-3.5	CNN/DM
1	0.73	0.75	0.79	0.56
2	0.55	0.56	0.60	0.47
3	0.45	0.45	0.48	0.43
4	0.42	0.41	0.47	0.41
5	0.39	0.35	0.42	0.39
6	0.34	0.33	0.37	0.40
7	0.31	0.32	0.38	0.39
8	0.30	0.29	0.34	0.36
9	0.27	0.24	0.31	0.37
10	0.24	0.24	0.28	0.35

Table 27: The average cosine similarity between reference summary and the first ten sentences in the article.