Fine-Grained Natural Language Inference Based Faithfulness Evaluation for Diverse Summarisation Tasks

Huajian Zhang[*] **Yumo Xu**[†] **Laura Perez-Beltrachini**
ILCC, School of Informatics
University of Edinburgh
huajian.zhang.21@gmail.com, {yumo.xu,lperez}@ed.ac.uk

Abstract

We study existing approaches to leverage off-the-shelf Natural Language Inference (NLI) models for the evaluation of summary faithfulness and argue that these are sub-optimal due to the granularity level considered for premises and hypotheses. That is, the smaller content unit considered as hypothesis is a sentence and premises are made up of a fixed number of document sentences. We propose a novel approach, namely INFUSE, that uses a variable premise size and simplifies summary sentences into shorter hypotheses. Departing from previous studies which focus on single short document summarisation, we analyse NLI based faithfulness evaluation for diverse summarisation tasks. We introduce DiverSumm, a new benchmark comprising long form summarisation (long documents and summaries) and diverse summarisation tasks (e.g., meeting and multi-document summarisation). In experiments, INFUSE obtains superior performance across the different summarisation tasks. [1]

1 Introduction

Current state-of-the-art summarisation systems are able to generate fluent summaries; however, their inability to generate factually consistent summaries remains a significant constraint in their real-world applications. As a result, the assessment of *summary faithfulness*, i.e., the degree to which a summary accurately represents the content of the input document, has recently received much research attention. This evaluation is key to assess progress in abstractive summarisation (Gehrmann et al., 2021, 2023). Existing research focuses on developing models to detect unfaithful summary content (Kryscinski et al. 2020; Scialom et al. 2021; Ribeiro et al. 2022; *inter alia*) as well as the meta-evaluation of these models with better benchmarks (Chen et al., 2021; Honovich et al., 2022; Durmus et al., 2022).

One way increasingly adopted to assess summary faithfulness is to use off-the-shelf Natural Language Inference (NLI; MacCartney and Manning 2009) models to determine whether a summary is entailed by the source document. NLI models determine the semantic relationship between a pair of texts: the *premise* and *hypothesis*. If the hypothesis can be inferred from the premise, it is said to be entailed by the premise. However, existing NLI models are mainly trained on relatively short texts from existing datasets Bowman et al. 2015; Williams et al. 2018. Examples in these datasets often represent inference cases over a single content unit (e.g., the example at the bottom of Figure 1 where inference is about the *transmission event*). This raises the question of how to apply them to produce entailment judgements for document-summary pairs consisting of multiple sentences aggregating several content units (e.g., the summary sentence MSS in Figure 1 aggregates content about the *company launching a legal action*, *a strike event*, and *the consequences of the strike*). Producing an entailment judgement for a summary sentence with several content units is a more complex entailment reasoning task.

Taking summary sentences as hypotheses, existing approaches try to either identify a document sentence that acts as the premise leading to the highest possible entailment score (sentence-level NLI, (Laban et al., 2022; Nie et al., 2020)) or directly measure entailment by taking the entire document as premise (document-level NLI, (Maynez et al., 2020; Honovich et al., 2022; Dziri et al., 2022)). However, due to content aggregation happening in summarisation, one document sentence will not contain enough content to entail a summary sentence. In Figure 1, none of the docu-

[*] Part of the work done for his MSc thesis at the University of Edinburgh.

[†] Work done while at the University of Edinburgh.

[1] Our code and data are available at https://github.com/HJZnlp/infuse

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1701–1722
March 17-22, 2024 ©2024 Association for Computational Linguistics

	D	$\models$ MSS	$\models$ SS$_1$	$\models$ SS$_2$
1	Lufthansa lost an appeal to a Frankfurt labour court, but is making a further legal challenge that could go late into Tuesday evening.	0.37	7.81	0.06
2	The pilots' strike, called over a pay dispute, will affect around 100,000 passengers, Lufthansa said.	0.61	0.69	1.74
3	The industrial action is part of a long-running pay dispute at Lufthansa.	0.18	0.74	0.06
4	The pilots' union Vereinigung Cockpit (VC) has organised 14 strikes since April 2014.	0.07	0.11	0.10
5	Short and medium-haul flights from Germany will be affected from 00:01 to 23:59 local time (23:01-22:59 GMT).	0.09	0.14	0.06
6	Flights by Lufthansa's other airlines including Eurowings, Swiss, Austrian Airlines, Air Dolomiti and Brussels Airlines are not affected by the strike, the airline said.	0.11	0.22	0.12
7	Pay talks between the Vereinigung union and the German airline broke down earlier this month, and Lufthansa said the union had "consistently rejected the offer" of mediation.	0.20	0.41	0.06
8	The union is calling for a 3.7% pay rise for 5,400 pilots dating back to 2012.	0.14	0.42	0.05
9	Lufthansa, which is facing increasing competition from budget rivals, offered a 2.5% increase over the six years until 2019.	0.12	0.22	0.11
10	Meanwhile, a separate dispute with cabin crew at Lufthansa's low-cost subsidiary, Eurowings, led it to cancel more than 60 flights on Tuesday.	0.27	0.47	0.32
MSS	German airline Lufthansa has launched a fresh legal challenge against a strike by its pilots, which could lead to the cancellation of more than 1,000 flights.			
SS	German airline Lufthansa has launched a fresh legal challenge against a strike by its pilots. The strike could lead to the cancellation of more than 1,000 flights.			

At 8:34, the Boston Center controller received a third transmission from American 11	$\models$	The Boston Center controller got a third transmission from American 11.

Figure 1: Example of input Document (D) and Model-generated Summary Sentence (MSS) from the AggreFact (Tang et al., 2023) benchmark on the XSum (Narayan et al., 2018) dataset. The example is considered unfaithful by the annotators. Simplified Summary (SS) is the generated summary after automatic sentence splitting. The cyan coulored text spans in the input document highlight those document content units that support the corresponding cyan spans in the summary. Red spans in the summary indicate content that is not supported by the input document. The $\models$ MSS and $\models$ SS$_i$ columns show entailment scores assigned by an off-the-shelf NLI model to document sentences acting as premises and either MSS or SS$_i$ sentences as hypotheses. The table in the bottom shows an example of entailment relation from the MNLI dataset (Williams et al., 2018). Entailment scores are computed by the NLI model introduced in Section 4 and normalised for better reading.

ment sentences alone can entail the complex summary sentence MSS aggregating several content units. On the other hand, taking the entire document as premise will perform poorly on long input documents (Schuster et al., 2022)). Recent work achieves promising results by first selecting an entailing context (context-level NLI, (Nie et al., 2019; Schuster et al., 2022; Kamoi et al., 2023)). That is, borrowing insights from information retrieval, these approaches carry out an initial step of document sentence retrieval to build a short context; and then perform NLI with the retrieved context as a premise. Specifically, in the retrieval step, given a summary sentence as hypothesis, document sentences are individually scored by an NLI model and ranked and the top k thereof constitute the premise (e.g., for the MSS in Figure 1, the 2^{nd}, 1^{st}, and 10^{th} would be selected as premise if $k = 3$).

In this work we argue that existing NLI-based approaches do not operate at the right level of granularity (Nenkova et al., 2007); even context-level NLI approaches. Summary sentences may convey several content units (Nenkova et al., 2007) partly overlapping with different document sentences. This renders the retrieval step of document sentences based on NLI scores less accurate (e.g., each document sentence in Figure 1 weakly entails the complex summary sentence MSS). In addition, summary sentences may aggregate content from different numbers of document sentences which makes it less accurate to have an entailing context with a fixed k number of document sentences (e.g., in Figure 1, SS$_1$ is entailed by two document sentences while SS$_2$ requires only one document sentence to show that its content is not derived from the document).[2] Finally, a fine-grained assessment of summary faithfulness brings interpretability, which hugely facilitates manual inspection of model-generated summaries.

We propose INFUSE, a faithfulness evaluation approach that **IN**crementally reasons over a document so as to arrive at a **F**aithf**U**lnes**S** **E**stimation of its summary. It aims at retrieving the best possible context to assess the faithfulness of each summary sentence (and in turn the entire summary), i.e., a context with the minimal and most relevant set of document sentences. Our incremental reasoning approach approximates this via successive expansions of the context adding document sentences and evaluating whether the hypothesis is entailed by it. Our approach further decomposes summary sentences for their faithfulness analysis. It does

[2] Note that *more than 1,000 flights* is not supported by the explicit facts stated in the input document.

this via sentence simplification. That is, it splits long summary sentences (e.g., MSS sentence in Figure 1) into a set of shorter ones conveying the same content units (e.g., SS_1 and SS_1 in Figure 1).

Most of previous work focuses on the meta-evaluation of NLI-based approaches on single document news summarisation (Laban et al., 2022; Tang et al., 2023). Thus, the question of how NLI-based evaluation works on diverse summarisation tasks is left unanswered. Hence, to widen the spectrum of NLI-based meta-evaluation (Gehrmann et al., 2021), we analyse the performance of NLI-based faithfulness evaluation approaches on long document summarisation with diverse domains and genres (Cohan et al., 2018; Huang et al., 2021; Zhong et al., 2021; Adams et al., 2023) and multi-document summarisation (Fabbri et al., 2019). We collect human annotated model-generated summaries from previous work on these tasks (Koh et al., 2022; Adams et al., 2023; Chen et al., 2023). We call this new set the DiverSumm benchmark.

We study existing NLI-based approaches on AggreFact (Tang et al., 2023), a benchmark for the meta-evaluation of single document summarisation, and DiverSumm. INFUSE achieves the best performance in these benchmarks. We find that the choice of an adequate level of granularity for the premise and hypothesis leads to more accurate entailment judgements when using off-the-shelf NLI models. On summaries of extractive nature, retrieving a small relevant set of document sentences suffices. Moreover, our results show that this is crucial for summarisation tasks with long input documents. Summary sentence splitting helps to obtain better performance in all summarisation tasks.

2 Faithfulness Annotated Data for Different Summarisation Tasks

Following previous work, we study faithfulness evaluation on two single document summarisation tasks, namely CNNDM (Nallapati et al., 2016) and XSum (Narayan et al., 2018). For this, we take the latest introduced faithfulness benchmark, AggreFact (Tang et al., 2023). It consists of a collection of document and model-generated summary pairs where summaries are annotated with faithfulness judgements by human judges. That is, each example in the benchmark is a triple (document, generated-summary, faithful/unfaithful label). AggreFact includes five annotated sets from the earlier SummaC (Laban et al., 2022) benchmark. These are XSumFaith (Maynez et al., 2020), FactCC (Kryscinski et al., 2020), SummEval (Fabbri et al., 2021), FRANK (Pagnoni et al., 2021), and Polytope (Huang et al., 2020). In addition, AggreFact includes four sets, namely QAGS (Wang et al., 2020) (referred as Wang'20 in the benchmark), CLIFF (Cao and Wang, 2021a), GOYAL'21 (Goyal and Durrett, 2021) and CAO'22 (Cao et al., 2022). AggreFact organises the annotated data into two major sets per summarisation task, CNNDM and XSum, herein we name them CNNDM_{AG} and XSum_{AG}. See Appendix A for details on the faithfulness annotation scheme of each dataset and the standarisation criteria applied to derive AggreFact.

DiverSumm a New Benchmark To study the performance of NLI-based faithfulness evaluation on diverse summarisation tasks, we propose a new benchmark, namely DiverSumm. It incorporates model generated summaries with human annotations about faithfulness from previous work (Koh et al., 2022; Adams et al., 2023; Chen et al., 2023). We follow (Laban et al., 2022) to standardise summary annotations into faithful/unfaithful labels. We discuss the summarisation task and characteristics of the annotated sets below.

ChemSumm (Adams et al., 2023) embodies the task of scientific long-form summarisation in the chemistry domain. Derived from academic journals, each input document contains section headers and associated paragraphs for all sections from the introduction up to the conclusion, and abstracts constitute the reference summaries.

MultiNews (Fabbri et al., 2019) is a large-scale multi-document news summarisation dataset with the number of input documents per example ranging from 2 to 6 and reference summaries written by editors.

QMSUM (Zhong et al., 2021) is a query-based multi-domain meeting summarisation dataset. It consists of meeting transcripts and queries associated with their corresponding abstractive summaries.

ArXiv (Cohan et al., 2018) is a long scientific paper summarisation dataset collected from ArXiv covering a wide range of topics. The main content up to the conclusion section of a paper is regarded as the document and the corresponding abstract section as the summary.

Summarisation Task	Doc.Tok	Sum.Sent	Sum.Tok	Cov	Dens	Summarisers
XSum (Tang et al., 2023)	360.54	1.01	20.09	0.55	0.99	OLD-EXFORMER, T5, BART, PEGASUS
CNNDM (Tang et al., 2023)	518.85	2.72	52.21	0.80	10.40	OLD-EXFORMER, T5, BART, PEGASUS
ChemSumm (Adams et al., 2023)	4612.40	7.36	172.79	0.91	10.89	LongT5, PRIMERA
QMSUM (Zhong et al., 2021)	1138.73	3.04	65.22	0.69	5.13	GPT-3.5, UniSumm, PEGASUS
ArXiv (Cohan et al., 2018)	4406.99	6.18	149.70	0.89	9.59	PEGASUS, BART
GovReport (Huang et al., 2021)	2008.16	15.07	391.22	0.86	12.76	PEGASUS, BART
MultiNews (Fabbri et al., 2019)	669.20	6.81	152.20	0.82	14.19	GPT-3.5, UniSumm, PEGASUS

Table 1: Statistics on AggreFact (test split) and DiverSumm per summarisation task. Document length in average number of tokens (Doc.Tok), summary length in average number of sentences (Sum.Sent) and tokens (Sum.Tok), and extractive metrics (Grusky et al., 2018) Density (Dens) and Coverage (Cov). Models generating summaries are LongT5 (Guo et al., 2022), PRIMERA (Xiao et al., 2022), GPT-3.5 (text-davinci-002) (Brown et al., 2020a; Ouyang et al., 2022), UniSumm (Chen et al., 2023), PEGASUS (Zhang et al., 2020), BART (Lewis et al., 2020), and T5 (Raffel et al., 2020). As grouped by Tang et al. (2023), OLD-EXFORMER denotes older models (See et al., 2017; Gehrmann et al., 2018; Liu and Lapata, 2019; Radford et al., 2019) .

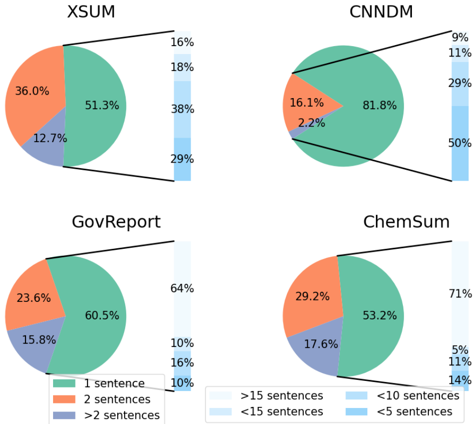


Figure 2: Statistics for the number of fused document sentences (the pie charts) and their distances (the blue vertical bars) on XSum and CNNDM (AggreFact) and GovReport and ChemSum (DiverSumm).

> **GovReport** (Huang et al., 2021) pairs long reports from government research agencies, including the Congressional Research Service and U.S. Government Accountability Office, with expert-written abstractive summaries.

Each summary in QMSUM and MultiNews was labeled using a 5-point Likert scale in terms of fluency, coherence, consistency, and relevance (Chen et al., 2023). We use the consistency criterion and label summaries as faithful if the score in consistency is 5, otherwise unfaithful. In ChemSumm, arXiv, and GovReport, summaries are annotated with a numerical number between 0 (inconsistent) and 1 (consistent) (Koh et al., 2022; Adams et al., 2023). We take summaries as faithful if the majority of the annotators labeled the summary as 1.

DiverSumm contains 563 test instances with a total of 4686 summary sentences of which 3138 have sentence level annotations. Table 1 shows relevant statistics about the benchmarks. Documents and summaries are longer in DiverSumm. Generated summaries for XSum and QMSUM are more abstractive (i.e., smaller coverage and density).

Error types Some subsets in AggreFact and DiverSumm, namely FRANK, ArXiv, and GovReport, contain sentence level and detailed error annotations for unfaithful summaries.[3] We exploit these annotations to analyse the performance of both the studied approaches and the NLI model on detecting different types of faithfulness errors. Concretely, unfaithful summaries are annotated with the following error types (Pagnoni et al., 2021). Relation Error (*PreE*) is when the predicate in a summary sentence is inconsistent with respect to the document. Entity Error (*EntE*) is when the primary arguments of the predicate are incorrect. Circumstance Error (*CircE*) is when the predicate's circumstantial information (i.e., name or time) is wrong. Co-reference error (*CorefE*) is when there is a pronoun or reference with an incorrect or non-existing antecedent. Discourse Link Error (*LinkE*) is when multiple sentences are incorrectly linked. Out of Article Error (*OutE*) is when the piece of summary contains information not present in the document. Grammatical Error(*GramE*) indicates the existence of unreadable sentences due to grammatical errors.

2.1 The Value of Adequate Premise and Hypothesis Granularity

We analyse document-summary pairs in the AggreFact and DiverSumm benchmarks to uncover the rational of why adequate premise and hypothesis granularity brings value into the evaluation of

[3] After manual inspection of the human annotations, we filtered out some examples in ArXiv and GovReport with a mismatch between the sentence and summary level annotation.

summary faithfulness (Nie et al., 2019; Schuster et al., 2022; Kamoi et al., 2023).

We examine the number of document sentences aggregated into a summary sentence via a greedy selection algorithm that maximizes document-summary token overlap (Lebanoff et al., 2019). As shown in Figure 2, 18-48% of summary sentences fuse more than one document sentence and at least 50% of the cases are not within a 5-sentence window. In particular, in GovReport 64% and ChemSumm 71% of the times the fused document sentences are in a 15 sentences or more window size. This renders sentence- and paragraph-level premises not ideal due to low recall. We show sentence fusion statistics for the other datasets in Figure 4, Appendix A.

An alternative to improve recall would be via increasing premise size. However, NLI models are typically trained on short premise-hypothesis examples with a premise average length ranging on 16-80 tokens for widely used datasets and a hypotheses length of 9-19 (Schuster et al., 2022). It is challenging for such models to generalise to document-level premises (average length is 439 in AggreFact and 2566 in DiverSumm). Previous work has shown that the performance of faithfulness evaluation consistently drops with longer premises (Schuster et al., 2022). We next describe our approach with premises of variable size (i.e, variable number of document sentences) and shorter hypotheses (i.e., simplified summary sentences).

3 InFusE

We denote a document as $D = \{d_m\}_{m=1}^{M}$ and a summary as $S = \{s_n\}_{n=1}^{N}$ where d_m and s_n are sentences. For a given summary sentence s_n as the *hypothesis*, we aim to retrieve a related context $R^{(n)}$, $R^{(n)} \subseteq D$, to act as the *premise* and estimate whether s_n can be entailed by $R^{(n)}$ (and, therefore, D) according to an NLI model θ. We assume that θ predicts one of the $\{\text{entailment}, \text{neutral}, \text{contradict}\}$ labels for a given premise-hypothesis pair. Summary sentence faithfulness estimates, given by $\theta(\text{entailment}|\cdot)$, are then aggregated into summary faithfulness scores with mean pooling.

Incremental Reasoning Given an NLI model θ, we construct a matrix E of entailment scores via sentence level inference between document sentences d_m and each summary sentence s_n. We derive from E entailment ranked lists of document sentences $\hat{D}^{(n)}$ associated to each summary sentence s_n. We then incrementally select sentences from $\hat{D}^{(n)}$ in a top-down fashion to retrieve a context $R^{(n)}$ for s_n. Starting from an empty context $R_0^{(n)}$, at each step i, we remove the top sentence from $\hat{D}^{(n)}$ and incorporate it into the current context to obtain a new context $R_i^{(n)}$. We then stop adding sentences to the context when the local minimum of the neutral class probability, $u_{i,n} = \theta(\text{neutral}|R_i^{(n)}, s_n)$, is reached, i.e., $u_{i,n} \geq u_{i-1,n}$. Intuitively, decreasing neutral scores signal shifts in the perceived entailment relationship from context $R_{i-1}^{(n)}$ to $R_i^{(n)}$ (i.e., candidate premises) and s_n (the hypothesis) leaning towards either entailment or contradiction. We stop when there is an increase in the neutral score. At this stopping point, the entailment score between the premise given by context $R_i^{(n)}$ and summary sentence s_n as hypothesis is taken as the final faithfulness estimation for s_n.

Algorithm 1 Summary sentence entailment estimation in InFusE.

```
Input: NLI model $\theta$, pair $(D, s_n)$.
Output: $R_{i-1}^{(n)}, e_{i-1,n}$ premise and entailment score for $s_n$.
for $d_m$ in $D$ do
    $e_{m,n}, u_{m,n}, c_{m,n} = \theta(d_m, s_n)$
    $e_{n,m}, u_{n,m}, c_{n,m} = \theta(s_n, d_m)$
    # entailment $e$, neutral $n$, contradiction $c$
    $\hat{E}_{d_m,s_n} = e_{m,n} + e_{n,m}$
end for
$\hat{D}^{(n)} = rank(\hat{E}_{d1:M}, s_n)$
$R_0^{(n)} = \emptyset, n_{0,n} = 0$
for $\hat{d}_i$ in $\hat{D}^{(n)}$ do
    add $\hat{d}_i$ to $R_i^{(n)}$
    $e_{i,n}, u_{i,n}, c_{i,n} = \theta(R_i^{(n)}, s_n)$
    if $u_{i,n} \geq u_{i-1,n}$ then
        stop and return $R_{i-1}^{(n)}, e_{i-1,n}$
    end if
end for
```

Reversed Reasoning In some cases, the content expressed in a document sentence d_m will only entail part of a summary sentence s_n (see example in Table 8 -bottom- of the Appendix). Thus, such d_m will have a low sentence level entailment score in E despite d_m really providing evidence for a part of s_n. Because summaries will contain extracted document fragments or paraphrases thereof, one way to improve entailment scores for such document sentences d_m is to reverse the direction in which sentence level NLI is applied. That is, we take the summary sentence s_n as premise and the document sentence d_m as hypothesis. We add reversed entailment scores to those on E and obtain a new re-weighted matrix $\hat{E}$ which is adopted to perform incremental context retrieval. Algorithm 1 summarises InFusE steps to estimate the entailment score of a given summary sentence with respect to its corresponding input document.

Sub-sentence Reasoning Different document sentences d_m will entail different parts of a summary sentence s_n (see document sentence fusion in Figure 2). In addition, those document sentences d_m may contain irrelevant content for s_n. Thus, sentence level scores in E as well as final context level entailment scores for s_n will be noisy (i.e., more chances of having neutral class high scores). Shorter summary sentences with finer-grained content units will yield more accurate contexts and entailment estimations. Figure 1 illustrates the difference in entailment scores in E when computed on the original summary sentence (MSS) and when computed on its sub-sentences (SS_1 and SS_2). In this work, we propose to simplify each summary sentence by splitting it into multiple sub-sentences.

4 Experimental Setup

We study NLI-based faithfulness evaluation approaches on AggreFact (Tang et al., 2023) and DiverSumm (Section 2). We adopt an ALBERT-xlarge (Lan et al., 2020) model optimized on MNLI (Williams et al., 2018) and VitaminC (Schuster et al., 2021) as our NLI model θ. MNLI covers ten distinct genres and styles of written and spoken English data. It aims to support a broader understanding and analysis of NLI across different genres and domains. VitaminC is synthetically created from Wikipedia sentences. Claim sentences are associated with contrastive evidence, i.e., one sentence that supports the claim and another one that does not. On MNLI (VitaminC) premises are 13.23 (43.03) tokens long in average and hypotheses 13.23 (27.57).

We fine-tune a T5-large (Raffel et al., 2020) model for sentence splitting. We use this model to simplify sentences in model generated summaries. We manually inspect several samples of split sentences and find that the performance is reasonable. Details about our sentence splitting model, examples, and statistics about the percentage of sentence splits are presented in Appendix B.

We compare InFusE with a widely adopted approach which considers the entire document as a premise, we refer to it as FullDoc. In practice, it divides the input document into chunks that fit the input size of the NLI model, computes chunks scores and takes the average thereof. We also compare with $\text{SummaC}_{\text{ZS}}$ (Laban et al., 2022), a sentence-level method which assumes each summary sentence is supported by one document sentence, and takes the one with the top entailment score as context. $\text{SummaC}_{\text{Conv}}$ introduces a convolutional layer trained on a subset of FactCC (Kryscinski et al., 2020) to aggregate the score given by an NLI model to each $\{\text{entailment}, \text{neutral}, \text{contradict}\}$ label into a final score. For a fair comparison with the other models, we remove specific constraints used in the original implementation of SummaC variants (see Appendix B). SenTLI (Schuster et al., 2022) retrieves a context with a fixed number k of document sentences. Its context includes document sentences with top entailment and top contradiction scores. Following (Schuster et al., 2022), we set the value of $k = 5$. We show performance with other values of k in Figure 6 in Appendix E. $\text{InFusE}_{\text{SUB}}$ is our variant with sub-sentence reasoning (i.e., summary sentence simplification). For this variant, to better mimic the process of label standardisation as described in Section 2, we use the *min(·)* operator to aggregate the entailment scores from sub-sentences into a sentence score.

5 Results

5.1 Faithfulness Evaluation

Following Laban et al. (2022), we adopt ROC-AUC (Bradley, 1997) which depicts classification performance with varied thresholds as our evaluation metric. Results on AggreFact and DiverSumm are shown in Table 2.[4] InFusE and $\text{InFusE}_{\text{SUB}}$ exhibit superior performance than previous approaches overall summarisation tasks. FullDoc exhibits the lowest performance, this confirms results from previous meta-evaluations (Laban et al., 2022; Schuster et al., 2022) and extends the observations to the summarisation tasks in our

[4] To determine the statistical significance of performance differences, we randomly re-sample 70% of the test instances 100 times and evaluate the models on these sets. We use the pairwise t-test to assess whether there is a significant difference between models.

	XSM$_{AG}$	CND$_{AG}$	CSM	QMS	AXV	GOV	MNW	AVG
FullDoc	72.77	64.40	50.15	37.12	62.78	79.19	44.76	58.74
SummaC$_{CONV}$	67.76	72.14	53.14	51.13	61.22	65.34	53.05	60.54
SummaC$_{ZS}$	70.29	74.54	54.41	48.21	69.44	79.37	50.17	63.78
SeNtLI	**73.61**	75.83	50.13	47.56	64.49	79.68	46.61	62.56
InFusE	73.42	**76.21**	54.11	52.16	71.38	**80.45**	**53.16**	**65.84**
InFusE$_{SUB}$	73.21	73.34	**59.26**	**53.20**	**73.89**	80.05	49.37	**66.05**

Table 2: Results for all summarisation tasks in AggreFact and DiverSumm. For AggreFact, we report the average results for XSum (XSM; 5 datasets) and CNN/DM (CND; 7 datasets), respectively; dataset-level performance can be found in Appendix D. CSM, MNW, QMS, AXV, and GOR refer to ChemSum, MultiNews, QMSUM, ArXiv, and GovReport respectively. We highlight **highest** scores and scores significantly different from FullDoc, SummaC variants, and SeNtLI approaches (at $p < .05$).

DiverSumm benchmark. SummaC$_{\text{CONV}}$, trained on specific evaluation data, does not generalise well across the different summarisation tasks. Thus, our main comparison variant is SummaC$_{\text{ZS}}$.

As for the role of sub-sentence reasoning, we observe that InFusE$_{\text{SUB}}$ works better on ChemSumm, QMSUM, and ArXiv where summary sentences are complex and informative (see sentence fusion in Figure 2) and more abstract (Table 1). This further validates the positive findings from claim verification tasks (Kamoi et al., 2023) for text summarisation. On the other hand, sub-sentence reasoning is less effective on CNNDM$_{AG}$, GovReport and MultiNews which consist of more extractive summaries (Table 1). In CNNDM$_{AG}$ segmenting short sentences may only introduce noise. This partially supports Glover et al. (2022) who draw a negative conclusion on the effectiveness of sub-sentence evaluation based on CNNDM. We note that the nature of the data underlying evaluation benchmarks should be further emphasized to delimit the scope of conclusions drawn. For GovReport and MultiNews, with the most extractive summaries (Table 1), we found that after splitting the relation between summary sub-sentences and document sentences becomes mostly one-to-one and thus approaches taking one document sentence become more effective (see results for existing approaches with sub-sentence hypotheses in Appendix C).

On XSum$_{AG}$, retrieval approaches, InFusE and SeNtLI, work very closely to the document-level approach FullDoc. The success of the document-level approach lies on the fact that summaries in XSum$_{AG}$ are highly abstractive (Table 1) and require reasoning over multiple document sentences; and input documents are short. Indeed, for highly abstractive summarisation tasks such as XSum$_{AG}$ or QMSUM it would make sense to build a structured premise with document sub-sentence content, connecting discourse information, explicit world knowledge, and intermediate inferences made explicit (Dalvi et al., 2021).

Results in Table 2 show that the variable premise size of InFusE leads to better performance across the board. In Appendix E, we show performance curves for InFusE versus InFusE$-k$, a version with different fixed premise sizes, to further illustrate this. We report statistics about the number of document sentences retrieved by InFusE and InFusE$_{\text{SUB}}$ in Table 4. These show the inherent variability in document sentence fusion happening within summary sentences (see approximation of this in Figure 2).

The reversed entailment direction in the retrieval step acts as a re-weighting scheme that takes advantage of paraphrased content and favors shorter document sentences (i.e., fewer content units than those appearing in summary sentences). We provide performance curves on the effect of reversed reasoning comparing InFusE with a version InFusE$-reversed$ in Appendix E; and case studies in Appendix G .

On DiverSumm, ROC-AUC scores are considerably lower than those obtained on AggreFact across the board. We attribute this to the summarisation tasks been more complex and recent models that generated the summaries more powerful. The lowest scores are on ChemSumm and QMSUM, we attribute this to the shift in vocabulary and genre in these tasks. The following lowest scoring task is MultiNews, we attribute this to the redundancy found in multi-document input. The fixed context of SeNtLI will only include redundant sentences.

5.2 Performance on Different Error Types

We look into the performance of the studied NLI-based approaches with respect to unfaithfulness errors discussed in Section 2. We focus on ArXiv, GovReport, and FRANK which contain fine-grained error annotations at sentence level. We consider each summary sentence in these subsets to be labelled with the error types that the majority of annotators agreed upon. We analyse the distribution of entailment scores for FULLDOC, SENTLI, INFUSE, and INFUSE$_{\text{SUB}}$ on summary sentences (i.e., without aggregation into a final summary score). We show these in Figure 3.[5]

Before looking into specific error types we analyse the range of scores the approaches assign to faithful sentences. Figure 3.a shows that FULLDOC tends to predict rather low entailment scores, close to zero, for most of faithful cases. This explains the lower ROC-AUC in Table 2. Using the entire document leverages noise when computing entailment if the input documents are long. After all, NLI models are trained to use the entire premise to yield a judgement and not to distinguish those relevant from irrelevant premise parts. Context-level approaches produce higher scores. INFUSE and INFUSE$_{\text{SUB}}$ produce more extreme scores.

On EntE error types, Figure 3.f shows that INFUSE assigns close to zero scores to more EntE cases. It works slightly better than INFUSE$_{\text{SUB}}$ and we attribute this to the fact that INFUSE$_{\text{SUB}}$ may introduce some noise when splitting sentences. In contrast, SENTLI assigns entailment scores in the range of [0.4,0.6] where also many faithful cases fall on. Figure 7.a/.b in Appendix F shows a similar trend for PredE and CircE error types.

As for discourse level errors, on LinkE error types, Figure 3.c, INFUSE works better. After manual inspection, we attribute this to the fact that none of the incorrectly fused document sentences contributes to a high entailment score. Thus, they will not be retrieved as an entailing premise. On CorefE errors, Figure 3.d, we can see that all approaches have poor performance assigning relatively high entailment scores. Note that the set of these error types is rather small.

Finally, on the OutE error type, Figure 3.e, INFUSE is better over INFUSE$_{\text{SUB}}$ and SENTLI. We attribute this to the fact that in cases of this error type there is no document information that can support nor contradict the summary sentence; thus, INFUSE will take the minimum number of document sentences (potentially only one) failing to entail the summary with OutE. Grammar errors, GramE in Figure 3.f, seem difficult to detect, which makes sense for NLI-based approaches.

We observed that in some error types INFUSE (and INFUSE$_{\text{SUB}}$) assigns extremely high (~ 1) scores to some cases. We manually examine a sample thereof and find that in most cases summary sentences have a high lexical overlap with document sentences and vary either on few tokens or word order. Thus, the NLI model is biased to rely on extractive cues (McKenna et al., 2023; Verma et al., 2023). Table 10 in Appendix F shows examples of error types correctly (~ 0) and incorrectly (~ 1) evaluated by INFUSE .

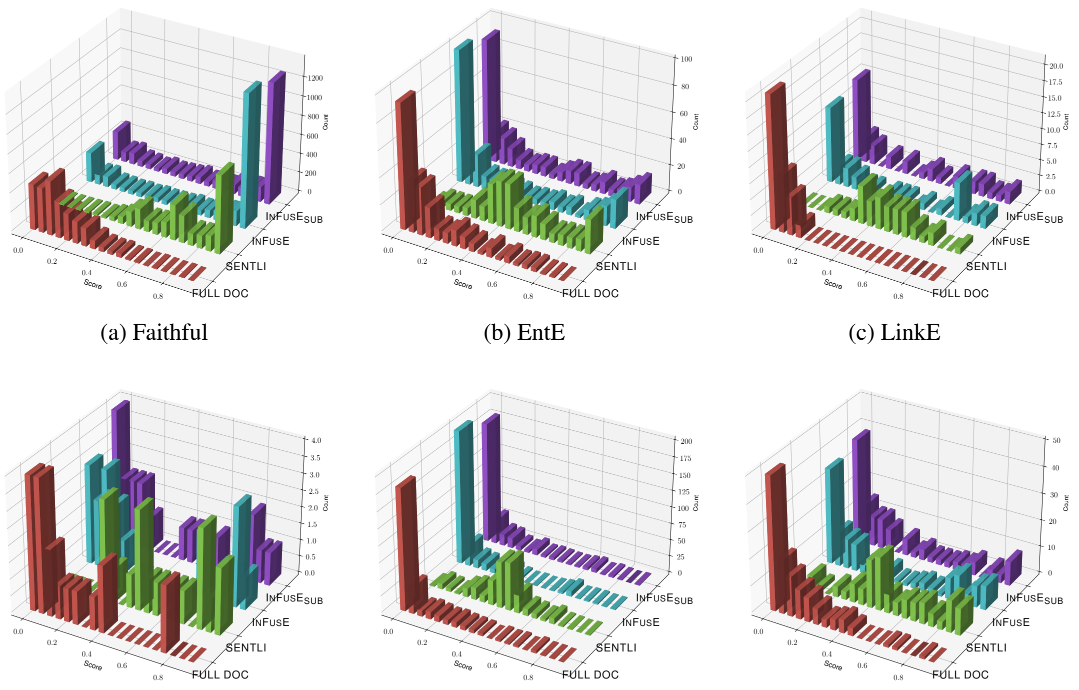
6 Related work

Some NLI-based approaches directly train document level NLI models (Yin et al., 2021; Utama et al., 2022). Others leverage off-the-shelf NLI models (Nie et al., 2019, 2020; Laban et al., 2022; Schuster et al., 2022; Kamoi et al., 2023; Steen et al., 2023). The former requires the construction of synthetic training data. In this paper, we study the latter type of approaches. These do not require additional data nor training resources.

(Nie et al., 2020; Laban et al., 2022) select a single sentence as premise while (Nie et al., 2019; Schuster et al., 2022; Kamoi et al., 2023) select a fixed number of document sentences, the same for all summary sentences. Our approach selects a variable number of document sentences as premise for each summary sentence. Recently, (Chen and Eger, 2023) conduct an empirical analysis of how to use the three directions in which entailment can be computed (entailment direction implication, reverse implication, and bi-implication). However, (Chen and Eger, 2023) directly use the scores from these directions in a single pass using the entire document as premise. In contrast, we apply reversed reasoning only to re-weight document sentences in the context retrieval step. (Steen et al., 2023) propose a document-level approach with data augmentation to adapt NLI models to task specific scenarios such as dialogue. Furthermore, they ensemble a number of calls to the NLI model via Monte-Carlo dropout to cope with domain shift. These ideas are orthogonal to our work and would make sense to use them in combination.

The value of fine-grained assessment of summary content has been highlighted in earlier work on summarisation evaluation (Marcu, 2001; Voorhees, 2004; van Halteren and Teufel, 2003; Teufel and van Halteren, 2004; Nenkova et al.,

[5] Note that for none of the approaches, we have tuned a faithful/unfaithful decision threshold; however, we compare faithful/unfaithful distributions and analyse performance at extreme scores 1/0 in the [0,1] interval.



(a) Faithful

(b) EntE

(c) LinkE

(d) CorefE

(e) OutE

(f) GramE

Figure 3: Distribution of entailment scores on faithful summary sentences and unfaithful ones encompassing different error types for ArXiv, GovReport and FRANK sets. The x-axe corresponds to the NLI-based approach. That is, FullDoc in red, SentLI in green, InFusE in cyan, and InFusE$_{\text{SUB}}$ in purple. The y-axe corresponds to the entailment scores (i.e., values ranging in [0,1]), and the z-axe corresponds to the count of instances.

2007; Gao et al., 2019; Shapira et al., 2019). This research highlights that summary sentences aggregate several content units and judgements should be initially provided for these before yielding a conclusion at summary level. However its focus is on the evaluation of content relevance. Recent work in the context of summary faithfulness evaluation assesses faithfulness of summary predicates and arguments (Goyal and Durrett, 2021). Conciliating with our results, they also show that fine-grained evaluation is beneficial. However, their approach is not based on NLI; and requires syntactic analysis of summary sentences and task specific human annotated data to train a classifier. Our approach is more generic and builds on existing resources. Contemporary with our work, (Min et al., 2023) propose the evaluation of Large Language Models (LLMs) generated biographies via their decomposition into smaller content units (i.e., atomic facts). Their approach is applied to factual descriptive generation. In contrast, we evaluate hallucination detection in a variety of summarisation tasks. For long dialogue summarisation, (Lattimer et al., 2023) propose to decompose the input into chunks, InFusE could be combined with a coarse chunk selection step.

Finer-grained evaluation has also shown positive results in the related task of claim verification (Chen et al., 2022; Kamoi et al., 2023). However, in the same way as current factuality evaluation on LLM generated text (Min et al., 2023; Manakul et al., 2023), they address more open-ended generation tasks where no ground truth input is assumed; their information source is either retrieved or parametric. We focus on NLI-based faithfulness evaluation from given input documents.

7 Conclusions

We study existing NLI-based faithfulness evaluation approaches and propose a new one, InFusE, that works at finer-grained granularity levels for computing document-summary entailment judgements. Our study shows that lower granularity via premises with variable size and summary sentence splitting is key to achieve more accurate entailment judgements when using off-the-shelf NLI models. We also introduce a new benchmark for long form input and diverse summarisation tasks. Experimental results show that InFusE achieves superior performance on evaluating faithfulness for diverse summarisation tasks.

Limitations

InFusE's stopping criteria can fall into a local minimum. In Table 4 (see Appendix E), we show the average number of document sentences retrieved by InFusE. It is evident that InFusE incremental context retrieval extracts more document sentences on XSum$_{AG}$ than on CNNDM$_{AG}$. This aligns with our analysis in Section 2.1 and the fact that summaries in XSum$_{AG}$ are more abstractive than those in CNNDM$_{AG}$. However, it might still not be enough, especially in XSum$_{AG}$, where some summary sentences indeed require more document sentences to form an entailing context. As a result, InFusE performance is comparable to SeNtLI which manually sets a fixed number of document sentences to be retrieved. This limitation can be overcome by introducing a hyper-parameter to the stopping criterion (Section 3); for example, to stop expanding the context when the neutral probability increases only by a large margin. The stopping criterion we adopt is simple but enough to show that it is possible to improve faithfulness evaluation performance when using off-the-shelf NLI models by allowing a variable premise size.

Another limitation of InFusE is that it requires additional calls to the NLI model. In Table 5, we show for all the compared approaches the computation cost of evaluating one summary sentence versus the achieved average performance. The reversed reasoning re-weighting in InFusE doubles the computation cost when compared with SeNtLI and SummaC. However, in practice, it would be possible to decrease the number of calls by using some heuristics to flag when it is necessary (or not). For instance, when the entailment score is above some threshold the reversed direction is not analysed; or the decision could be based on whether the summary sentence fuses more than one document sentence which can be computed based on a cheap metric such as ROUGE. The automatic stopping criteria requires a number of additional calls given by the expected number of retrieved document sentences k_{avg} taken as premise. The complexity of inference for a context-level approach with a fixed number of retrieved sentences k, i.e., InFusE-k or SeNtLI, assuming a standard transformer, is $O(k^2)$ whereas for InFusE it is in $O(k_{avg}^3)$. If k_{avg} is small enough and there is variability in the number of retrieved document sentences, which is the case in the analysed summarisation tasks (see Table 4 in Appendix E), InFusE can be competitive in terms of running times. Summary sentence splitting also adds an extra overhead; however, it will decrease summary sentence fusion of document sentences, i.e., fewer cases will need reversed reasoning and k_{avg} will be smaller. In terms of performance, the contribution of reversed reasoning and dynamic stopping can be seen in Figure 6 in Appendix E. Although grid search for k will give the best possible k, this k value will be the same for all summary sentences (within a summary and within a dataset). In contrast, dynamic stopping lets each summary sentence be analysed with a different k value. Figure 6 shows that InFusE with dynamic stopping is better than InFusE-k for different values of k.

Acknowledgements

We thank our reviewers for their constructive feedback. We are grateful to Griffin Adams, Huan Koh, Yulong Chen, and Yang Liu for making available the annotated datasets that we include in DiverSumm. We also thank Mirella Lapata for useful feedback. We gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1).

References

Griffin Adams, Bichlien Nguyen, Jake Smith, Yingce Xia, Shufang Xie, Anna Ostropolets, Budhaditya Deb, Yuan-Jyue Chen, Tristan Naumann, and Noémie Elhadad. 2023. What are the desired characteristics of calibration sets? identifying correlates on long form scientific summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10520–10542, Toronto, Canada. Association for Computational Linguistics.

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

Andrew P Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens

Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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. 2021a. CLIFF: contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6633–6649. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021b. 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.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yanran Chen and Steffen Eger. 2023. MENLI: Robust Evaluation Metrics from Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Yiran Chen, Pengfei Liu, and Xipeng Qiu. 2021. Are factuality checkers reliable? adversarial meta-evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2082–2095, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2023. UniSumm and SummZoo: Unified model and diverse benchmark for few-shot summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12833–12855, Toronto, Canada. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rodolfo Delmonte, Antonella Bristot, Marco Aldo Piccolino Boniforti, and Sara Tonelli. 2007. Entailment and anaphora resolution in RTE3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 48–53, Prague. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. 2022. Spurious correlations in reference-free evaluation of text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1443–1454, Dublin, Ireland. Association for Computational Linguistics.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

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.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir

Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Yanjun Gao, Chen Sun, and Rebecca J. Passonneau. 2019. Automated pyramid summarization evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 404–418, Hong Kong, China. Association for Computational Linguistics.

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

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, 77.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

John Glover, Federico Fancellu, Vasudevan Jagannathan, Matthew R. Gormley, and Thomas Schaaf. 2022. Revisiting text decomposition methods for NLI-based factuality scoring of summaries. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–105, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

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, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

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

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How far are we from robust long abstractive summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682–2698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Barrett Martin Lattimer, Patrick Chen, Xinyuan Zhang, and Yi Yang. 2023. Fast and accurate factual inconsistency detection over long documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1691–1703. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

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

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

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.

Daniel Marcu. 2001. Discourse-based summarization in duc-2001.

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

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2758–2774. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

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

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2):4–es.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6859–6866.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4812–4829. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching sentence-pair NLI models to reason over long documents and clusters. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 394–412, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

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.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.

Julius Steen, Juri Opitz, Anette Frank, and Katja Markert. 2023. With a little push, NLI models can robustly and efficiently predict faithfulness. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 914–924, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Simone Teufel and Hans van Halteren. 2004. Evaluating information content by factoid analysis: Human annotation and stability. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 419–426, Barcelona, Spain. Association for Computational Linguistics.

Prasetya Utama, Joshua Bambrick, Nafise Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776, Seattle, United States. Association for Computational Linguistics.

Hans van Halteren and Simone Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 57–64.

Dhruv Verma, Yash Kumar Lal, Shreyashee Sinha, Benjamin Van Durme, and Adam Poliak. 2023. Evaluating paraphrastic robustness in textual entailment models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 880–892, Toronto, Canada. Association for Computational Linguistics.

Ellen Voorhees. 2004. Overview of the trec 2003 question answering track.

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

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

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

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

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

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

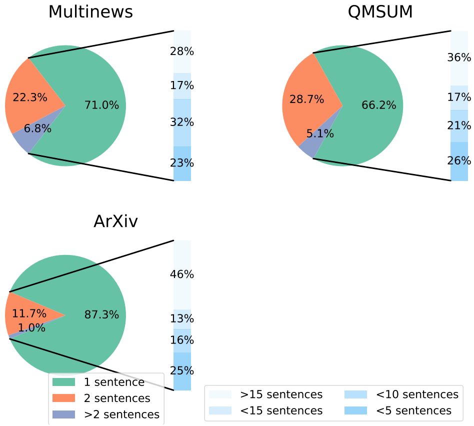


Figure 4: Statistics for the number of fused document sentences (the pie charts) and their distances (the blue vertical bars) on qmsum, multinews, and arxiv (DiverSumm).

A Additional Dataset Details

Annotated Sets in AggreFact FactCC by Kryscinski et al. (2020) and SummEval Fabbri et al. (2021) are annotated at summary level. FactCC uses a binary consistency label (consistent/inconsistent). SummEval uses a 5-point Likert scale where only a score of 5 is treated as correct while the rest are considered incorrect.

The annotation of Wang'20 Wang et al. (2020) and FRANK Pagnoni et al. (2021) operates at sentence level. Wang'20 employs a binary consistency label (consistent/inconsistent). A summary is labelled as faithful (consistent) if all of its sentences are labelled as consistent. The annotation scheme in FRANK highlights faithfulness error types (see Error Types in Section 2) in summary sentences. Summaries in FRANK are considered to be faithful if none of their sentences are annotated with errors.

Polytope (Huang et al., 2020), XSumFaith Maynez et al. (2020) and Goyal'21 Goyal and Durrett (2021) are annotated at span level. Polytope identifies various error types such as addition, omission, and intrinsic inaccuracies. The annotation of XSumFaith revolves around error types like intrinsic and extrinsic. Goyal'21 classifies the error types into intrinsic, extrinsic × entity, event, noun phrase. Summaries devoid of these errors are marked as faithful.

CLIFF Cao and Wang (2021b) is annotated at word level and its annotation scheme accounts for instinct/extrinsic hallucinations and lack of world knowledge. Cao'22 by Cao et al. (2022) annotates entities and categorizes incorrect entities into factual/non-factual/instinct hallucinations. Sum-

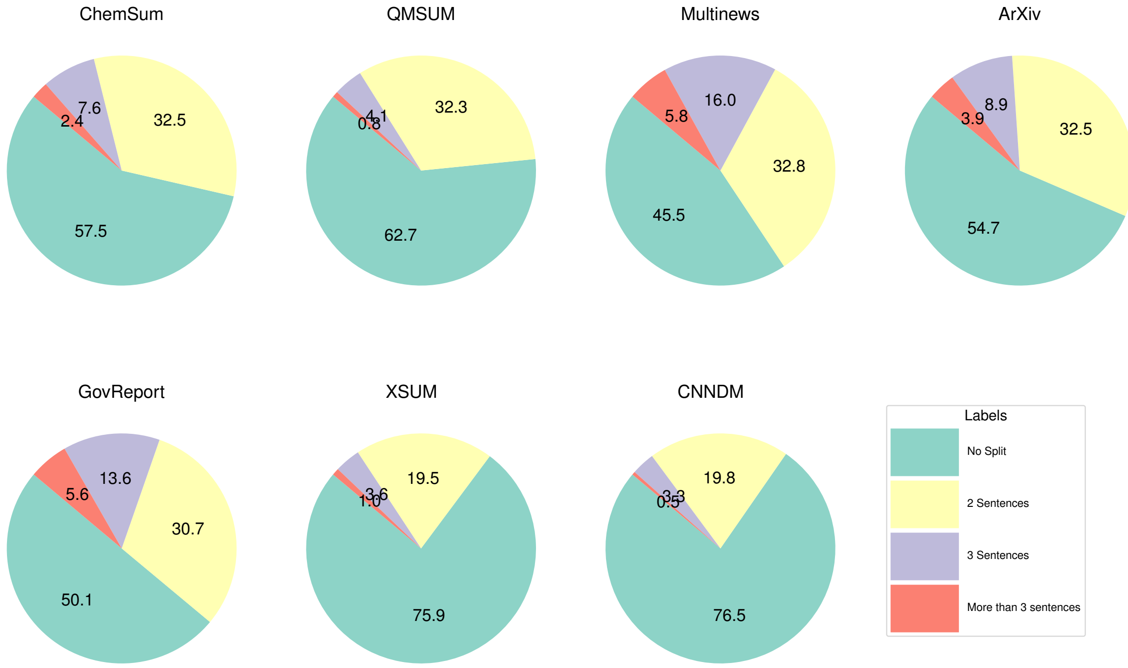


Figure 5: Distribution of number of splits occurring in summary sentences.

Sentence	Sub-sentences
Heritage auctions offered the gray jacket featuring a black zigzag applique	Heritage auctions offered the gray jacket. The gray jacket featured a black zigzag applique.
S.t. Mirren have signed striker Jeremy Clarkson on a season-long loan from Dundee.	S.t. Mirren have signed striker Jeremy Clarkson. The striker is on a season-long loan from Dundee.
Change is a problem for many disabled people.	Change is a problem for many disabled people.

Table 3: Examples of original sentences and their rewritten sentences for sub-sentence reasoning.

maries devoid of these errors are marked as faithful.

For details on the annotation process, we refer the reader to Aggrefact (Tang et al., 2023).

License No license is found for AggreFact, GovReport and ChemSumm. ArXiv is under Apache-2.0 license. QMSUM and MultiNews are under MIT License. We ensure that the data was used solely for academic purposes, which aligns with the intended use of these datasets. For data safety, content filtering was conducted when the creators built the original datasets. It is not avoidable that some documents can contain uncomfortable content, including news coverage of crimes and wars. For the model-generated summaries annotated with human judgements collected from (Chen et al., 2023; Koh et al., 2022; Adams et al., 2023) to create DiverSumm, we download some sets from their corresponding online download link and make others directly facilitated by the authors available in our github.[6] We obtained permission from the authors for their use and encourage citation of the sets' corresponding work upon their future use within DiverSumm. We use these annotated sets only for research purposes.

B Training Configurations

Models We use the publicly-available `https://huggingface.co/tals/albert-xlarge-vitaminc-mnli` NLI model. We use the tokenizer from Stanza (Qi et al., 2020).

[6] `https://huggingface.co/datasets/griffin/ChemSum`, https://github.com/huankoh/How-Far-are-We-from-Robust-Long-Abstractive-Summarization.

Models	XSM$_{\text{AG}}$	CND$_{\text{AG}}$	CSM	QMS	AXV	GOV	MNW
INFUSE	2.66±1.67	1.79±1.04	2.50 ±4.76	2.55 ±1.48	4.89 ±9.25	2.11±1.21	1.98 ±1.25
INFUSE$_{\text{SUB}}$	2.40±1.57	1.22±1.75	2.12 ±3.79	2.41 ±1.43	4.09 ±8.26	1.92 ±1.11	1.76 ±1.09

Table 4: Average number of retrieved document sentences and standard deviation for INFUSE and INFUSE$_{\text{SUB}}$ on AggreFact and DiverSumm.

Approach	AUC	Nb. calls to NLI
FULLDOC	58.74	1
SUMMAC$_{\text{CONV}}$	60.54	M+C
SUMMAC$_{\text{ZS}}$	63.78	M
SENTLI	62.56	M+1
INFUSE-k	65.01	2M+1
INFUSE	65.84	2M+k_{avg}+1

Table 5: Performance / Computation trade-off. We report the AUC versus the number of calls to the NLI model. M is the number of document sentences. k_{avg} is the expected number of retrieved document sentences which can entail the summary sentence. C is the call to the convolution layer.

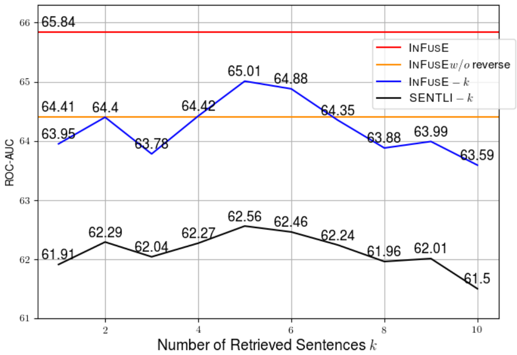
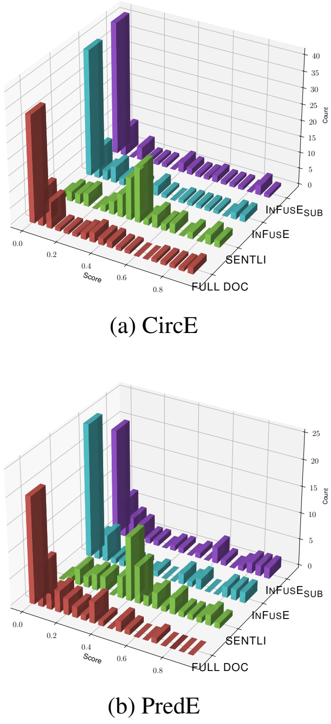


Figure 6: Performance over retrieval size k. We report the average ROC-AUC on AggreFact and DiverSumm.



(a) CircE

(b) PredE

Figure 7: Distribution of entailment scores on correct and different error types for arXiv and GovReport from DiverSumm. The x-axe corresponds to the NLI-based approach. That is, FULLDOC in red, SENTLI in green, INFUSE in cyan, and INFUSE$_{\text{SUB}}$ in purple. The y-axe corresponds to the entailment scores (i.e., values ranging in [0,1]), and the z-axe corresponds to the count of instances.

Originally SUMMAC$_{\text{ZS}}$ uses the combination of entailment - contradiction which was found to perform better (Laban et al., 2022). However, we find that in both AggreFact and DiverSumm, by taking only the entailment score SUMMAC$_{\text{ZS}}$ obtains a much better performance. Thus, we only use entailment scores. In addition, the implementation of SUMMAC ignores those document sentences with less than 10 tokens and only considers the first 100 sentences of the document. We remove such constraints for a fair comparison. In addition, SUMMAC obtains better performance without these constraints.

Sentence Splitting Kamoi et al. (2023) propose a dataset, namely WiCE, including original claim sentences paired with their decomposition (split) into more than one sentence generated by GPT-3 (Brown et al., 2020b). We leverage such parallel data to train a sentence splitting model for sub-sentence reasoning based on T5-large (Raffel et al., 2020). We fine-tune T5-large for 5 epochs with a batch size of 32 and a learning rate 5e-4. We force the length of the output to be within $[3, 128]$. We show a few sentence splitting examples in Table 3. Figure 5 shows the distribution of the number of splits that summary sentences had. We train the model on an A6000 GPU and each epoch costs 90 seconds. The inference time is around 8 sample per second.

	XSM$_{AG}$		CND$_{AG}$		CSM		QMS		AXV		GOV		MNW		AVG	
CONTEXT	SENT	SUB	SENT	SUB	SENT	SUB	SENT	SUB	SENT	SUB	SENT	SUB	SENT	SUB	SENT	SUB
FULLDOC	72.77	**73.63**	64.40	63.68	50.15	**58.72**	37.12	**39.76**	62.78	62.46	79.19	77.69	44.76	**46.72**	58.74	**60.38**
SUMMAC$_{CONV}$	67.76	65.77	72.14	70.84	53.14	51.10	51.13	**54.42**	61.22	44.26	65.34	**81.58**	53.05	**56.27**	60.54	**60.61**
SUMMAC$_{ZS}$	70.29	66.67	74.54	**74.98**	54.41	**57.32**	48.21	**51.42**	69.44	67.26	79.37	**81.09**	50.17	**54.20**	63.78	**64.71**
SENTLI	**73.61**	71.45	75.83	74.66	50.13	**55.69**	47.56	**51.88**	64.49	**76.35**	79.68	77.65	46.61	43.61	62.56	**64.47**
INFUSE	73.42	73.21	**76.21**	73.34	54.11	**59.26**	52.16	**53.20**	71.38	**73.89**	**80.45**	80.05	**53.16**	49.37	**65.84**	**66.05**

Table 6: Results for all summarisation tasks in AggreFact and DiverSumm combined with summary sentence splitting (SUB column). For AggreFact, we report the average results for XSum (XSM; 5 datasets) and CNN/DM (CND; 7 datasets), respectively; dataset-level performance can be found in Appendix D. CSM, MNW, QMS, AXV, and GOV refer to ChemSumm, MultiNews, QMSUM, ArXiv, and GovReport respectively. We highlight **highest** scores and scores significantly different from FULLDOC, all SUMMAC variants and SENTLI models (at $p < .05$, using pairwise t-test). We additionally highlight in **olive** improved scores for existing approaches when combined with summary sentence splitting.

C Summary Sentence Splitting is Beneficial for All Approaches

Table 6 shows additional results when we combine the proposed summary sentence splitting step with the different approaches to build a premise. We can see that sub-sentence (SUB column in Table 6) brings improvements across all of them (as discussed before with the exception of CNN/DM). Sub-sentence evaluation brings improvements for sentence-level premises such as SUMMAC in particular for the version that relies on a convolutional neural network trained to map the distribution of entailment scores to correct/incorrect judgements. After splitting there is less content fusion from document sentences and more feasible to judge entailment with one document sentence.

For ArXiv and CSM context-level works better indicating that neither one sentence nor the entire document provide adequate context even after summary sentence splitting. For XSum, the most abstractive dataset with short input documents, the document-level (FULLDOC) and context-level (INFUSE and SENTLI) premises work well. For this dataset sentence-level approaches (SUMMAC) even with sentence splitting are not enough. Overall, INFUSE and INFUSE$_{SUB}$ perform the best, this shows that the variable context allows to account for different levels of document sentence fusion.

D Dataset-Level Performance on AggreFact

We show detailed results for AggreFact in Table 7. Statistical significance of INFUSE w.r.t. to the other best performing approaches are computed as described in Section 5.1. Overall, there is no significant difference among INFUSE , SENTLI, and FULLDOC on XSum$_{AG}$ and CNNDM$_{AG}$. Interestingly, the models exhibit different performance within subsets of the tasks. INFUSE is significantly better on Wand'20, CLIFF, and FactCC.

E Performance per Premise Sizes

Figure 6 shows the evaluation performance (ROC-AUC) for different premise sizes k (i.e., number of document sentences). It includes SENTLI , a variant of INFUSE with a fixed retrieval size (INFUSE-k), and INFUSE without reverse reasoning. As can be seen, reversed reasoning helps to produce better entailment judgements as there is a performance degradation when we remove it from INFUSE. Incremental reasoning allows INFUSE to determine *when to stop* automatically, removing the requirement of additional data for optimizing the retrieval size k which has a substantial impact on model performance.

Table 4 shows the average premise size, in number of document sentences, at which INFUSE and INFUSE$_{SUB}$ work. We can see that there is considerable variability in the number of retrieved sentences within and across tasks. This further supports the difference in performance between INFUSE and INFUSE-k.

F Performance per Error Types

Figure 7 shows two additional graphs for CircE and PredE error types. Similarly to EntE (Section 5.2), INFUSE and INFUSE perform better than SENTLI which assigns scores mainly in the [0.4, 0.6] interval. INFUSE performs better than INFUSE$_{SUB}$. We show examples of cases correctly (~ 0) and incorrectly (~ 1) judged by INFUSE in Table 10.

Models	XSum Test						CNN/DM Test							
	Wang'20	Cao'22	XSF	Goyal'21	CLF	**AVG**	FCC	Wang'20	SEV	PTP	FRK	Goyal'21	CLF	**AVG**
FullDoc	64.62	**71.50**	**75.76**	74.70	77.34	72.78	75.63	**84.09**	74.07	76.69	65.26	4.17	70.90	64.40
SummaC$_{\text{Conv}}$	69.59	69.71	70.03	56.40	73.09	67.76	92.22	76.67	85.48	81.67	76.72	**25.00**	67.20	72.14
SummaC$_{\text{ZS}}$	73.77	67.27	72.73	61.58	76.11	70.29	93.72	80.94	87.57	88.57	77.22	**25.00**	68.81	74.54
SeNtLI	72.80	70.57	69.46	**74.88**	80.34	**73.61**	92.26	80.04	87.75	92.82	**79.92**	20.83	**77.23**	75.83
InFusE	**76.41**	67.73	74.01	71.43	77.51	73.42	**94.99**	80.21	**88.65**	**92.84**	79.48	20.83	76.45	**76.21**
InFusE$_{\text{SUB}}$	73.76	69.92	74.69	66.34	**81.36**	73.21	92.73	78.66	87.76	83.68	77.76	16.67	72.82	72.87

Table 7: Dataset-level performance on AggreFact. For XSum Test, XSF and CLF refer to XSumFaith and CLIFF, respectively. PTP, FCC, SEV and FRK refer to Polytope, FactCC, SummEval and FRANK, respectively. We highlight **highest** scores and scores significantly different from FullDoc, all SummaC and SeNtLI models (at $p < .05$, using pairwise t-test).

G Case Studies

Sentence Fusion To illustrate how sentence fusion renders difficult the assessment of entailment by current sentence-level NLI models, we provide two representative examples from the faithfulness evaluation benchmarks in Table 8.

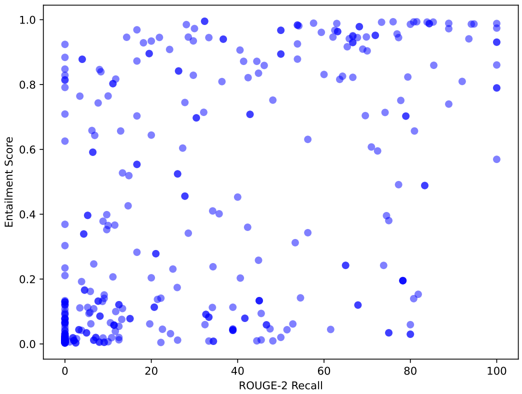


Figure 8: ROUGE-2 Recall versus Entailment Score on summary sentences labelled as unfaithful from the ArXiv and GovReport datasets.

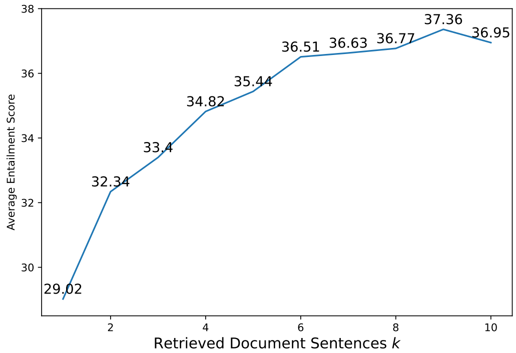


Figure 9: Average entailment score for summary sentences labelled as unfaithful from the ArXiv and GovReport datasets. Premise size, in number of retrieved document sentences, ranges from 1 to 10.

The first example, taken from FactCC (Kryscinski et al., 2020), shows a summary that is simply a short version of one document sentence. In these cases, a sentence-level NLI evaluator (Laban et al., 2022; Nie et al., 2020) would capture the relation and assign a high entailment score.

In contrast, the second example from CAO'22 (Cao et al., 2022) is more complex: the content conveyed in the summary sentence fuses content included in multiple document sentences. In this situation, three possibilities arise. First, if the summary sentence is more informative than a document sentence and the part that overlaps is a paraphrase, it can be captured by applying NLI in reversed direction (i.e., summary-to-document, MSS $\models$ DS column in Table 8). Examples of this scenario are text segments highlighted in cyan. Second, if none of the inference directions (neither document-to-summary, DS $\models$ MSS column in Table 8, nor MSS $\models$ DS) achieve a high entailment score individually, the combined score may still be relatively high allowing the bidirectional method in InFusE to capture such cases as illustrated by the example in green. Third, a content unit in a complex and informative summary sentence is entailed by a content unit in a complex document sentence they only overlap on this content unit. It is possible that the method will fail in these cases, as the sentence segments in violet illustrate.

High Reversed Reasoning Scores Table 9 shows examples of document and summary sentence inference applied in both the standard and reversed direction (lines 2 and 3 in Algorithm 1). These examples are taken from summaries annotated as (correct) faithful. In particular, these show cases where the reversed direction yields high entailment scores. These are cases where the summary sentence is pro-

		DS $\models$ MSS	MSS $\models$ DS
D	Sao Paulo, Brazil (CNN)Brazilian supermodel Gisele Bundchen sashayed down the catwalk at Sao Paulo Fashion Week on Wednesday night in an emotional farewell to the runway.	.003	.003
	Bundchen announced over the weekend that she would be retiring from the catwalk, though not the fashion industry.	.004	.003
	The 34-year-old, who is married to New England Patriots quarterback Tom Brady and has two children, has said she wants to spend more time with her family.	.001	.001
	On Wednesday night, Brady had a front-row seat at what was hailed as a historic moment in Brazil's fashion world.	.006	.003
	Bundchen wrote about her fashion career on her Instagram account: "I am grateful that at 14, I was given the opportunity to start this journey.	.996	.002
	Today after 20 years in the industry, it is a privilege to be doing my last fashion show by choice and yet still be working in other facets of the business."	.002	.001
MSS	bundchen wrote about her fashion career on her instagram account.		
D	David Lipton, second in command at the IMF, outlined some of these risks in a speech to the National Association for Business Economics in Washington on Tuesday.	.018	.001
	"The IMF's latest reading of the global economy shows once again a weakening baseline," he said.	.103	.006
	"We are clearly at a delicate juncture."	.020	.212
	The comments come after weaker-than-expected trade figures from China showing that exports plunged by a quarter from a year ago.	.004	.001
	The IMF has already said it is likely it will downgrade its current forecast of 3.4% for global growth when it next releases its economic predictions in April.	.050	.020
	The dismal picture is one that has on-going ramifications for businesses and industries that bet on China's growth story.	.002	.003
	Read more from Karishma:	.002	.004
	Why a story about bulk shipping matters.	.002	.019
MSS	The head of the International Monetary Fund (IMF) has warned that the global economy is "at a delicate juncture" and that the outlook for global growth is "deteriorating".		

Table 8: We show input Document (D), Model-generated Summary Sentence (MSS), and DS $\models$ MSS (Document Sentence -DS- to summary sentence reasoning) and MSS $\models$ DS (reversed reasoning) scores. We highlight content segments in summary sentences and their corresponding document evidence in violet, cyan and green. The example in the top part is from FactCC (Kryscinski et al., 2020) in CNNDM$_{\text{AG}}$ and the second is from CAO'22 (Cao et al., 2022) in XSUM$_{\text{AG}}$. Both labelled as faithful (correct) summaries.

viding more details due to sentence fusion. For instance, in the third example, the summary sentence is adding extra information (taken from other document sentences) about *Paulo Duarte* being *Burkina Faso's coach*. Note that in some cases sentences contain pronouns and thus they should not lead to high entailment scores because the referent is unknown (Delmonte et al., 2007). However, the NLI model is biased because of the premise-hypothesis length and token overlap (McKenna et al., 2023; Verma et al., 2023).

~ 0 and ~ 1 Entailment Scores on Different Error Types Table 10 shows examples of INFUSE working on FRANK, ArXiv, and GovReport (Section 5.2). The top part of the table includes cases where INFUSE successfully assigns close to zero scores to unfaithful cases per error type and the bottom part illustrates those scenarios where it fails to identify the error. On manual inspection, we find that in many cases these failures are related to high lexical overlap and premise-hypothesis length bias (McKenna et al., 2023; Verma et al., 2023). Figure 8 and Figure 9 show this trend for all unfaithful sentences in the ArXiv and GovReport subsets. We observe a similar trend in all datasets in AggreFact and DiverSumm but only these two datasets have sentence level annotation.

In Figure 8, we analyse entailment scores for premise-hypothesis pairs in relation to their lexical

DS	MSS	DS $\models$ MSS	MSS $\models$ DS
He resigned from his post in order to make this appearance.	A police chief resigned from his post to appear on bbc question time.	.003	.938
We will be making no appeal.	Wigan warriors will not appeal against the eight-game ban given to ben flower for punching st helens prop lance hohaia.	.004	.930
"I am confident they can recover in time," Duarte insisted.	Burkina faso coach paulo duarte says he is confident his players will be fit for next month's africa cup of nations.	.013	.388
In a statement the company said the blaze had affected an estimated 1,000-2,000 tonnes of recycled wood chip.	Firefighters are continuing to tackle a blaze at a wood chip recycling plant in Bridgend county which has destroyed up to 2,000 tonnes of wood chip.	.019	.067
Decisions about which people, groups, or events to memorialize are made by many different entities, including Congress, federal agencies, state and local governments, and private citizens, among others.	Decisions about which people, groups (or events), and which places to memorialize, are made by many different entities, including Congress, federal agencies, state and local governments, and private citizens, among others.	.091	.980
NOAA has defined natural infrastructure and nature-based infrastructure in NOAA Administrative Order (NAO) 216-117: NOAA National Habitat Policy.	NOAA's National Habitat Policy (NAO 216-117) directs the agency to protect, maintain, and restore ocean, coastal, and Great Lakes ecosystems by "applying natural and natural infrastructure," among other activities.	.007	.685
This report considers the extent of federal involvement in memorials located outside the District of Columbia (Washington, DC).	This report considers the extent of federal involvement in national memorials located outside the District of Columbia (Washington, DC).	.166	.981
In the United States, there are hundreds, and possibly thousands, of memorials to various individuals, groups, and events.	In the United States, there are hundreds, and possibly thousands, of memorials to various individuals, group, and historical events.	.224	.989

Table 9: Examples of reversed reasoning with high entailment scores. Document Sentence (DS), Model-generated Summmary Sentence (MSS), document to summary entailment (DS $\models$ MSS), and reverse direction (MSS $\models$ DS). All examples are from summaries in the DiverSumm benchmark labelled as faithful (correct).

overlap.[7] We compute lexical overlap as ROUGE-2 Recall in order to capture phrase information. As can be seen, on the left-bottom corner, a high percentage of pairs with low ROUGE-2 Recall obtain a low entailment score. Another cluster of pairs is on the right-top corner where pairs with high lexical overlap get high entailment scores. This behaviour of NLI models will undermine evaluation of summary faithfulness when summaries are abstract or have a high token overlap but differ in few words that change the content conveyed in the input document (e.g., negation). Figure 9 shows average entailment scores in relation to premise size. That is, we compute average entailment scores for premise-hypothesis pairs setting the premise to the top k ranked document sentences; k takes values from 1 to 10. We can see that longer premises obtain higher entailment scores despite the fact that they include document sentences further below in the rank.

[7] Note that by premise we mean the premise selected by INFUSE.

DRS	MSS	Error Type
Entailment Scores ∼ 0		
Costs for Group B benefits and administration are financed by the one-time appropriation of $4.6 billion provided in the Zadroga Reauthorization Act of 2015.	Costs for Group B benefits and administrative expenses were financed by a one-time appropriation of $3.	EntE
Jan 2006 - Government proposes nuclear as part of future energy mixMar 2013 - Construction of Hinkley Point approvedOct 2013 - UK government agrees £92.50 per megawatt-hour will be paid for electricity produced at the Somerset site - around double the current market rate at the timeOct 2015 - EDF signs investment agreement with China General Nuclear Power Corporation (CGN)July 2016 - EDF board approves final investment decision, but the UK Government postpones a final decision on the project until autumn.	The government has given the go-ahead for a new nuclear power plant at a former nuclear plant in somerset.	PredE
The VCF was reauthorized in 2015 and, if not reauthorized in the 116 th Congress, will sunset on December 18, 2020.	The MTF was reauthorized in 2015 and, if not reauthorized, the current iteration will sunset on June 18, 2017.	CircE
While men caregivers may face some of these risks, the effects of caregiving for women are compounded by lower average lifetime earnings and a longer life expectancy than men. As a result, women caregivers are at an increased risk of outliving their savings.	Women caregivers were more likely than men caregivers to be employed and to have higher levels of earnings, but women caregivers were also more likely to work part-time and have lower levels of employment and have less income.	LinkE
In our December 2018 report, we found that TSA provides pipeline operators with voluntary security guidelines that operators can implement to enhance the security of their pipeline facilities.	The Transportation Security Administration (TSAO) provides pipeline operators with voluntary security guidelines that operators can implement to enhance the security of their pipeline facilities.	CorefE
Since fiscal year 2008, the United States has allocated about $3 billion for assistance for Mexico under the Mérida Initiative. You asked us to review issues related to Mérida Initiative implementation and objectives.	Since fiscal year 2008, the United State has allocated about $3 billion for assistance for Mexico under the Civil Standards Initiative.	OutE
In July 2016, OMB issued an updated Circular No. A-123, Management's Responsibility for Enterprise Risk Management and Internal Control, which requires executive agencies to implement enterprise risk management (ERM) in their management practices. Since the July 2016 update to OMB Circular No. A-123 required agencies to implement ERM, the Air Force has been leveraging and relying on its existing risk management practices.	In July 2016, OMB issued an updated Circular No A. B, Management's Responsibility for Enterprise Risk Management and Internal Control, which requires executive [incomplete sentence]	GramE
Entailment Scores ∼ 1		
Practitioners and decisionmakers have been using the term nature-based infrastructure and supporting nature-based infrastructure features since at least the late 2000s (although these types of features have been assigned various names over time)	Practitioners and decisionmakers have been using the term nature-by-nature-infrastructure since at least the late 2000s, although these types have been assigned various names over time.	EntE
Memorials with "medium" federal involvement typically either are located on federal land but do not receive federal funding, or are located on nonfederal land but receive assistance from a federal agency.	Memorials for purposes of "medium" involvement are either located on nonfederal land but do not receive federal funding, or are located in federal land but receive federal assistance from a federal agency.	PredE
But he now faces at least a year at a militant rehabilitation centre in Kuwait, according to the terms of the release. The Kuwaiti government had pushed hard for the release of all Kuwaiti detainees at Guantanamo.	A former guantanamo bay detainee has been released from kuwait.	CircE
The value of the 15 State projects in our sample is about $88 million, and the value of the five USAID projects in our sample is about $107 million. Because State/INL implemented about 90 percent of Mérida Initiative projects during this period, we chose a larger State/INL sample than a USAID sample.	State/INL and USAID have implemented about 90 percent of MérIDA Initiative projects.	LinkE
Administrators of the ACT test took the decision just hours before some 5,500 students were due to sit it. The other entrance exam - the SAT - was cancelled in South Korea in 2013 because some of the questions were leaked.	A number of students have been barred from taking part in a test test test in south korea.	CorefE
But Prof Peter Godfrey-Smith said the unique study, based on 53 hours of footage and published on Friday in the journal Current Biology, provided a novel perspective on octopus behaviour."[An aggressive] octopus will turn very dark, stand in a way that accentuates its size and it will often seek to stand on a higher spot," Prof Godfrey-Smith, who co-authored the report, said.	One of the world's most aggressive octopuses appears to show signs of aggressive behaviour, a study suggests.	OutE
No systematic law or set of regulations governs the establishment of memorials outside Washington, DC.	No systematic law or set of regulations governs the establishment of memorialses outside Washington, D.C.	GramE

Table 10: Examples of unfaithful summaries per error type which correctly obtain low scores by InFusE (top block) and incorrectly high scores (bottom block). We indicate the document sentences retrieved by InFusE (DRS), the Model-generated Summary Sentence (MSS), and Error Type according to (Koh et al., 2022).