# Leveraging Digitized Newspapers to Collect Summarization Data in Low-Resource Languages

**Noam Dahan** **Omer Kidron** **Gabriel Stanovsky**

The Hebrew University of Jerusalem

noam.dahan1@mail.huji.ac.il

## Abstract

High quality summarization data remains scarce in under-represented languages. However, historical newspapers, made available through recent digitization efforts, offer an abundant source of untapped, naturally annotated data. In this work, we present a novel method for collecting naturally occurring summaries via *Front-Page Teasers*, where editors summarize full length articles. We show that this phenomenon is common across seven diverse languages and supports multi-document summarization. To scale data collection, we develop an automatic process, suited to varying linguistic resource levels. Finally, we apply this process to a Hebrew newspaper title, producing HebTeaseSum, the first dedicated multi-document summarization dataset in Hebrew[1].

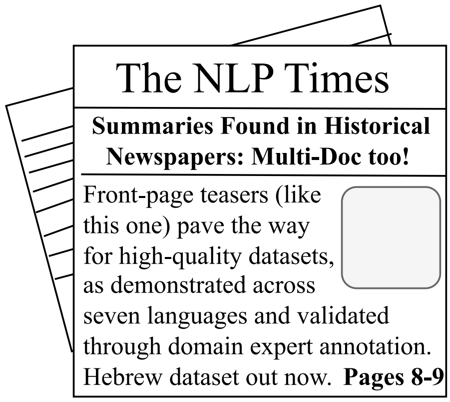


Figure 1: Newspaper's front-page teasers are a natural source for high-quality, expert-written summaries of full articles appearing inside the paper. We show that they are common across languages, and lend themselves to straightforward data collection.

## 1 Introduction

Recent studies suggest that the task of summarization in English may be already solved, or even "(almost) dead" (Pu et al., 2023; Zhang et al., 2024). However, this is not the case in the vast majority of world languages, which suffer from lack of accessible, high-quality summarization datasets (Dahan and Stanovsky, 2025). Collecting high-quality summarization data is hard: human annotation in summarization is particularly challenging (Varab and Schluter, 2021), while automatic data collection methods depend on the availability of web content, which is frequently limited in low-resource settings (Joshi et al., 2020).

In this work, we observe that print newspapers contain organic, high-quality summarization data. In particular, we leverage *front-page teasers*, short blurbs written by professional editors describing one or more articles appearing inside the issue (Utt and Pasternack, 1989), as illustrated in Figure 1. We will show that front-page teasers constitute high-quality, expert-written summaries, which are abundant in a range of languages.

We begin by demonstrating that front-page teasers can be collected in a simple, two-step process, and show that they are widely available by collecting data in seven diverse languages. Furthermore, we validate the suitability of teasers as summaries through expert annotation, which found them to be high quality and consistent with the source document. Interestingly, we find that front-page teasers are also a rich source for *multi-document summarization*, when a single teaser summarizes several news stories.

Following, we use front-page teasers to evaluate state-of-the-art LLMs on summarization across diverse languages. We find that LLMs struggle to fully cover all of the information in the gold summary. Moreover, performance disparities between models are significantly larger in lower-resource languages, underscoring the need to curate high-quality datasets in these settings.

Finally, we show that front-page teasers are also

[1] https://github.com/edahanoam/HebTeaseSum



*Findings of the Association for Computational Linguistics: EACL 2026*, pages 5260–5273
March 24-29, 2026 ©2026 Association for Computational Linguistics

I dag åpner NM på ski og første dagen kan bli rekruttenes dag. Elin Nilsen B&Y IL, og Inger Lise Hegge, Henning, er begge rekrutter som tidligere i sesongen har tuktet landslagsløperne. 10 kilometer er distansen kvinnene går i dag, mens mennene skal gjøre unna 30 kilometer. I tillegg er det kombinert hopprenn og laghopprenn.

Sporten side 4

Vonir um að mynda þjóðstjórn í Makedóníu dvínuðu mjög í gær, þar sem albanskur stjórnmálaflokkur krefst þess að árásum á vopnaða albanska íbúa í Kosovo verði hætt áður en viðræður hefjast.

bls. 2.

Alessandra Casella, 27 anni, è la coprotagonista di «Fred», in questi giorni al teatro Alfieri di Torino. Dopo l'esperienza all'Actor's Studio e molto teatro, vorrebbe lavorare di più per il cinema ma, ci ha detto (l'intervista è a pagina 27), «in Italia son tempi duri, i registi preferiscono le indossatrici».

(A pag. 27)

Figure 2: **Front-Page Teasers are common in a variety of languages**. Highlighted text shows the reference to relevant pages where the corresponding articles can be found, and serves as a useful signal for identifying summaries. From Left-to-Right: teasers from Rana Blad (Norway), Fréttablaðið (Iceland) and Stampa Sera (Italy). Links to newspapers and translations are in Appendix A.

a source for large scale evaluation or fine tuning. Using a simple heuristic over a large corpus of Hebrew print newspapers, we create HEBTEASESUM, the first dedicated summarization dataset for Hebrew which supports multi-document summarization, totaling 7,774 samples.

To conclude, we make the following contributions:

- We introduce a novel approach for building summarization datasets, using untapped naturally annotated data of historical newspapers.
- We show that our approach is applicable to many languages and supports model evaluation across varying linguistic resource levels.
- We release a multi-document summarization dataset in Hebrew, created using our proposed approach.

## 2 Background

In recent years, extensive efforts have been made to digitize historical newspapers, making them increasingly accessible for NLP research. In this section, we describe this growing resource and clarify the journalism terminology used in this work.

**Extensive Digitization Efforts.** National libraries worldwide preserve collections of historical newspapers in many languages (Beals et al., 2020). The U.S. Library of Congress alone curates texts in 36 languages, including native languages considered to be endangered (Zhang et al., 2022). Similar multilingual newspaper archives can be found across Europe as well as in national libraries of at least 163 countries (Haneefa and Jiji, 2019).

These newspapers are becoming increasingly accessible to researchers due to vast digitization and open access efforts. In 2015, it was estimated that there are more than 45 thousand digitized newspaper titles, while the actual number may be significantly higher (Kettunen et al., 2020). This is especially important for low resource languages, as many of them have minimal web presence (Joshi et al., 2020).

Moreover, recent advances in OCR models such as Mistral OCR and improved post-correction methods (Thomas et al., 2024) give hope for overcoming past challenges in using this valuable data, especially the problems caused by errors in text recognition and segmentation (Dell et al., 2023).

**Terminology.** We define the key terms we use in this work. *Newspaper titles* refer to publications such as "The New York Times" or "The Daily Mail", while an *issue* refers to a specific edition of a newspaper title, for example, the issue of The New York Times that came out on January 1st, 1990. We refer to the short text appearing on the front page and pointing to an article on a different page within the issue as a *front-page teaser*. These are also known as 'lead-in' or 'synopsis' (Utt and Pasternack, 1989). Front page teasers are different from *leads* of articles sometimes appearing on the front page, that contain the *beginning* of an article and point the reader to its continuation elsewhere in the issue.

## 3 Front-Page Teasers are Common, High-Quality Summaries

In this section, we describe our approach to extract a summarization dataset given a collection of printed newspapers. We conduct human evaluation performed by domain experts, to demonstrate front-page teasers high quality. Then, we show the organic front-page summaries heuristic holds across several languages with various levels of resources, and naturally supports multi-document summarization.

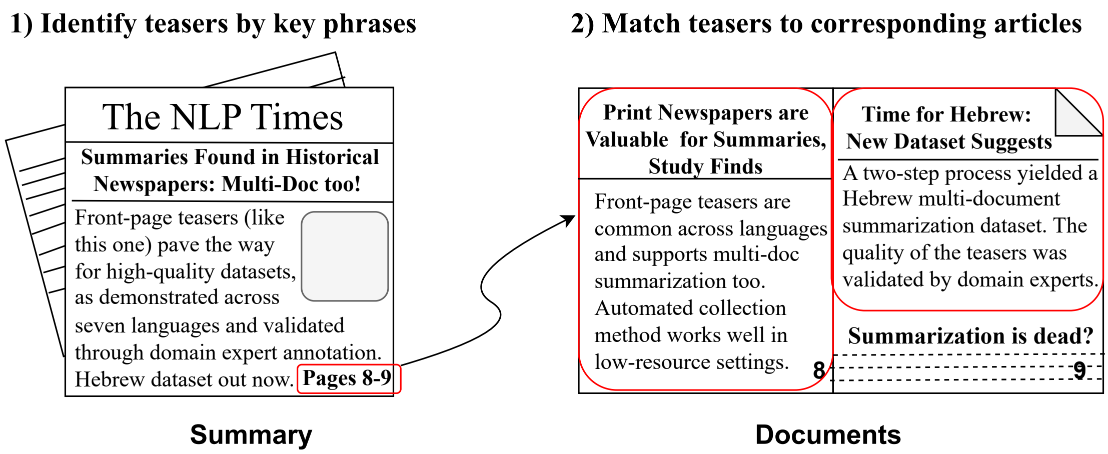


Figure 3: **Our approach for extracting a summarization dataset from printed newspapers.** We first find teasers and corresponding page numbers on the front page based on newspaper-specific keywords (e.g., "Full articles on Pages 8-9"). Then, we turn to these pages and find the articles most relevant to the content of the teaser, thus resulting in (teaser summary, relevant articles) pairs.

### 3.1 Constructing Summarization Data from Teasers

Newspapers often provide readers with front-page teasers, summarizing the main points of one or more articles in the issue, see Figure 2 for examples. These teasers have been popular for at least the last 50 years (Utt and Pasternack, 1989). Driving our work is the observation that front-page teasers are high-quality, multi-document summaries which can provide valuable signal for low resource languages.

To collect a dataset reflecting that observation, the process consists of two steps, as shown in Figure 3: First, identifying the teasers, and second, matching them to the articles to which they point.

Identifying front-page teasers is a simple process that requires determining the key terms each newspaper title uses to indicate that a full story is available in the issue. For example, teasers often end with a phrase like *"page 5"*, indicating where the full article appears.

The second step, after identifying the front-page teasers, is to turn to the relevant pages to locate the full articles related to it. This step is more challenging than the first, for several reasons: First, the teaser and the article often have different headlines. Second, the same page may contain articles on similar topics. Finally, we find that front-page teasers rarely mention how many articles the summary supports. Nevertheless, a careful reading allows readers to identify the relevant articles.

Following these two steps produces a collection of articles and the teasers summarizing them in an abstractive and organic manner.

### 3.2 Human Evaluation

To assess front-page teasers' suitability as summaries, we conduct human evaluation by two domain experts, each with over five years of experience in journalism. Their analysis, detailed in this section, showed that teasers are generally of high quality, with length serving as a useful indicator. Results are presented in Table 1.

We randomly select 400 samples from the Hebrew newspaper used in our dataset along with 836 corresponding articles. For a meaningful evaluation, we divide the teasers into four groups based on length and ensured an equal number of samples (100) from each category. A translated example for each of the categories is provided in Appendix B.

We follow Fabbri et al. (2021) to score each teaser according to four metrics: (1) *Coherence*, measuring how well the sentences in the summary fit together and whether they sound natural; (2) *Consistency*, measuring whether the facts in the summary are consistent with the facts in the article; (3) *Fluency*, measuring the quality of each sentence individually and (4) *Relevance*, measuring how much of the important information appears in the summary.[2] Results are presented in Table 1. To measure inter-annotator agreement, the annotators overlapped on 100 teaser annotations (25 per category), achieving a Krippendorff's alpha (Krippendorff, 2018) of 0.74, an agreement level often achieved in similar tasks (Antoine et al., 2014).

All length categories achieved high scores in

[2]Annotator guideline can be found in Appendix C

| Len. | Coherence | Consistency | Fluency | Relevance |
|---|---|---|---|---|
| 0-25 | 4.31 | 4.75 | 4.36 | 3.29 |
| 25-50 | 4.47 | 4.73 | 4.74 | 3.90 |
| 50-100 | 4.63 | 4.68 | 4.91 | 4.20 |
| >100 | 4.88 | 4.32 | 4.98 | 4.30 |

Table 1: **Average teaser scores provided by domain expert annotators.** "Len."' denotes teaser length (in words). Relevance increases with teaser length while other scores are generally high.

fluency and coherence, as they were written by professional editors to be featured on the front-page. The most common cause of score reductions in these metrics was the use of "bullet style" sentences, which, while grammatically correct, were perceived as less natural by our annotators.

Moreover, the teasers were found to be consistent with the source document. In all length spans we found that front-page teasers contain information that is present in the source articles. This serves as a valuable indicator of the quality of teasers as summaries, especially given that web sourced summaries from the news domain were previously found to contain information that does not appear in the articles they are based on (Tejaswin et al., 2021). The difference may lie in the fact that, while web-scraping heuristics extract parts of the article (such as the first sentence or the title) to be used as summaries, teasers are intended to be read independently from the news article as they appear on a different page.

The Relevance score exhibited the highest score variance, with length influencing how much of the important information is included in the teaser. In both the under-25-word and under-50-word categories, the main relevance issue raised by the annotators was that the summary omitted some information from the source, accounting for 70% of the explanations for scores below 4. This is an inherent challenge of very short summaries. The second most common issue in these length categories was that the teaser consisted primarily of a significant quote from the document (11% of the explanations given in scores below 4). Quotes often highlighted only a few key aspects rather than the full content. This was particularly problematic in the shortest teasers (under 25 words), where the quote could comprise all or most of the teaser.

### 3.3 Front-Page Teasers are Common in a Wide Range of Languages

We show that front-page teasers are a widespread phenomenon by collecting summary-article pairs in seven languages from diverse families as shown in Table 2.

We leverage digitization efforts to select a newspaper title for each language, ensuring we can access full scans and navigate between pages. We then collect summary-article pairs by manually matching front-page teasers to their corresponding articles, until we reach 30 summaries per language, following Shaib et al. (2024), which found that such sample size could be sufficient to compare model performance on news summarization.

Matching teasers to articles enables systematic data collection for a wide range of languages. On average, nine issues per title are needed to collect 30 samples, highlighting both the abundance of untapped, naturally annotated data and the manageable effort required to gather sufficient data. Moreover, all newspaper titles featured multi-document summarization, when the teaser summarized several articles. Appendix A details titles and dates.

In Table 2, we report the level of abstraction in the collected data using the common novel n-gram ratio (Narayan et al., 2018), which measures the percentage of n-grams that appear in the front-page teasers but not in the corresponding article. We also report the compression ratio, comparing the length of the teaser to that of the article. In the next section, we use this data to evaluate the performance of state-of-the-art LLMs in summarizing diverse languages against organic, challenging data.

## 4 Experiments

We use the manually collected front-page teasers data to evaluate the performance of state-of-the-art LLMs on news articles in seven languages. Below, we present the metrics we use to assess summary quality, then describe our use of LLM-as-a-judge to evaluate some of these aspects. We then report our findings, presented in Table 3.

### 4.1 Metrics and LLM-as-a-Judge

We report ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) to measure overlap between generated summaries and gold references, a practice that remains common in summarization evaluation (Gao et al., 2025; Li et al., 2024). ROUGE is limited to textual overlap so we additionally report

| Language | Res.* | Title | Comp.** | Novel n-grams | | | | %Multi-doc*** |
|---|---|---|---|---|---|---|---|---|
| | | | | 1 | 2 | 3 | 4 | |
| Norwegian† | 1 | Rana Blad | 0.51 | 0.36 | 0.64 | 0.75 | 0.80 | 13% |
| Icelandic | 2 | Fréttablaðið | 0.83 | 0.41 | 0.63 | 0.72 | 0.76 | 6% |
| Estonian | 3 | Eesti Päevaleht | 0.67 | 0.59 | 0.78 | 0.85 | 0.90 | 3% |
| Greek | 3 | Kathimerini | 0.61 | 0.56 | 0.85 | 0.94 | 0.97 | 16% |
| Hebrew | 3 | Hadashot | 0.84 | 0.49 | 0.74 | 0.84 | 0.89 | 26% |
| Italian | 4 | Stampa Sera | 0.79 | 0.36 | 0.60 | 0.69 | 0.74 | 3% |
| Polish | 4 | Dziennik Polski | 0.72 | 0.38 | 0.58 | 0.65 | 0.70 | 30% |

Table 2: **Characteristics of the summaries collected across languages.** *Res. indicates the language resource level, as categorized by (Joshi et al., 2020), showing that we cover languages with vßarying levels of resources. **Comp. stands for compression rate reflecting the document-to-summary length ratio (Grusky et al., 2018). ***%Multi-doc represents the portion of the summaries which span multiple documents. Front-page teasers represent high level abstractivity across metrics. †We use Norwegian Bokmål throughout this work.

BERTScore (Zhang et al., 2019) to capture semantic similarity.

To complement traditional metrics, which may miss key aspects of information overlap (Deutsch and Roth, 2021), we incorporate three additional evaluation metrics to complement them: *coherence* and *consistency*, which were also used in the human evaluation process, and *coverage*, which assesses the extent to which the predicted summary captures the information presented in the reference summary (Liu et al., 2022). These qualities can help distinguish between a model's ability to generate abstractive summaries in a given language (as measured by coherence and consistency) and its ability to identify the key information.

Ideally, we would have employed human evaluators to assess these properties, but this is infeasible due to the limited availability of annotators in low-resource languages (Bhat and Varma, 2023), therefore we rely on LLM-as-a-judge, an evaluation approach that has garnered considerable traction in recent years (Zheng et al., 2023). Hada et al. (2024) have recently established that in the context of summarization, evaluation using a strong model and a detailed prompt aligns with human annotations. However, we note that some works warn that LLM evaluation might be skewed in favor of model generated text (Forde et al., 2024), thus showing optimistic results, especially in low resource languages. To make the evaluation more reliable, we use a reference-based method. Additionally, we use GPT-4o solely as the evaluation judge and not as one of the summarization models.

### 4.2 Results

We evaluate three models: DeepSeek-R1 (Guo et al., 2025), Llama 3.3 70B (AI@Meta, 2024) and Mixtral-8x7b (Jiang et al., 2024) and report results on three shot settings. Temperature is set to 0 to ensure consistent and deterministic outputs. For LLM-as-a-judge we use GPT-4o (Hurst et al., 2024). We present results in Table 3. Below we discuss several observations.

**Generated summaries miss aspects of the gold summary.** We observe differences between consistency and coverage across all models and languages. This suggests that while the generated summaries are grounded in the source text (i.e., consistent), they often focus on different content than what appears in the front-page teasers (i.e., lower coverage). There is no substantial variation in performance in this aspect across models within the same language: the average difference in coverage between the best and worst performing models is only 0.345 out of 5. A similar pattern is observed in BERTScore, where the largest difference across models in the same language is just 0.04, suggesting that all models exhibit similar limitations in fully capturing the content of the gold summaries. One possible explanation is that front-page teasers differ from previously used news summaries, which are typically based on web articles and may present a new challenge for LLMs.

**There is a bigger difference in model performance in lower resource languages.** Most models seem to generate coherent text which is consistent with the source documents in all languages,

| Language | Model | ROUGE | | | BERTscore | LLM-as-a-Judge (1-5 scale) ↑ | | |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | L | | Coherence | Consistency | Coverage |
| Norwegian | Mixtral | 0.30 | 0.08 | 0.27 | 0.71 | 4.00 | 4.33 | 2.70 |
| | Llama | **0.33** | **0.11** | **0.30** | **0.72** | 4.41 | 4.7 | 2.70 |
| | DeepSeek-R1 | 0.26 | 0.06 | 0.24 | 0.70 | **4.52** | **4.78** | **2.74** |
| Icelandic | Mixtral | 0.19 | 0.04 | 0.17 | 0.68 | 3.08 | 3.38 | 2.23 |
| | Llama | **0.27** | **0.11** | **0.25** | **0.72** | 4.46 | 4.62 | **2.92** |
| | DeepSeek-R1 | 0.21 | 0.06 | 0.18 | 0.69 | **4.58** | **4.69** | 2.62 |
| Estonian | Mixtral | 0.09 | 0.02 | 0.08 | 0.63 | 2.65 | 2.65 | 2.7 |
| | Llama | **0.11** | **0.05** | **0.11** | 0.64 | 4.70 | 4.7 | 3.09 |
| | DeepSeek-R1 | 0.07 | 0.02 | 0.07 | 0.64 | **4.78** | **4.91** | **3.35** |
| Greek | Mixtral | 0.14 | 0.03 | 0.12 | 0.67 | 3.93 | 4.19 | 2.63 |
| | Llama | **0.16** | 0.04 | **0.14** | **0.68** | **4.48** | 4.48 | 2.74 |
| | DeepSeek-R1 | 0.15 | 0.04 | 0.13 | 0.67 | 4.44 | **4.74** | **3.00** |
| Hebrew | Mixtral | 0.14 | 0.04 | 0.13 | 0.68 | 2.86 | 2.93 | 1.93 |
| | Llama | **0.21** | **0.09** | **0.20** | **0.70** | 3.96 | 4.18 | **2.50** |
| | DeepSeek-R1 | 0.13 | 0.03 | 0.12 | 0.68 | **4.54** | **4.82** | 2.25 |
| Italian | Mixtral | 0.28 | 0.09 | 0.25 | 0.69 | 4.40 | 4.36 | **2.84** |
| | Llama | **0.30** | **0.10** | **0.28** | 0.69 | 4.48 | 4.64 | 2.76 |
| | DeepSeek-R1 | 0.24 | 0.05 | 0.21 | 0.68 | **4.64** | **4.88** | 2.80 |
| Polish | Mixtral | 0.21 | 0.07 | 0.20 | 0.68 | 4.12 | 4.08 | 2.32 |
| | Llama | **0.23** | **0.10** | **0.21** | 0.68 | 3.96 | 3.84 | 2.12 |
| | DeepSeek-R1 | 0.17 | 0.05 | 0.16 | 0.67 | **4.36** | **4.36** | **2.36** |

Table 3: **Model performance on our multilingual data collection in a three shot setting.** The best number per metric for each language is highlighted in bold. While some models produce *coherent* text which is *consistent* with the content of the article, all models struggle with fully *covering* the information in the gold summary.

with coherence and consistency scores exceeding 4.3 out of 5 and showing similar levels across all languages. However, we observe larger performance gaps between models on lower resourced languages. For example, in Icelandic, the gaps between the best and worst performing models in coherence and consistency are 1.5 and 1.3, respectively; for Estonian, the differences are 2.13 and 2.26; and for Hebrew, 1.68 and 1.89. In contrast, the highest-resource languages in our study, Italian and Polish, show much smaller gaps: 0.24 and 0.52 for Italian, and 0.4 and 0.52 for Polish. These findings highlight the need for high-quality data in lower-resource languages to enable meaningful model comparisons. Interestingly, Norwegian, the lowest-resource language in our work, scores highly across all metrics and models. Further experiments are needed to determine whether this reflects genuinely better performance or whether the metrics are less reliable for low-resource languages like Norwegian.

Although these results offer a preliminary indication of the front-page teasers potential to assess multilingual summarization capabilities, a more robust model evaluation and further training could require larger-scale datasets, beyond what can feasibly be collected through manual annotation. Thus, in the next section, we present an automated method for collecting such data at scale.

## 5 Automatic Extraction of Teaser Summaries and Corresponding Articles

We describe an automatic process for building a large-scale summarization dataset. We begin by detailing the automated two-step process: identifying front-page teasers and matching them to their corresponding articles. This is followed by an evaluation of the method, assessing its practicality and suitability for low-resource settings. Finally, we introduce our dataset, HEBTEASESUM, created by applying this approach to a single Hebrew newspaper title, yielding over 7,000 summaries, nearly half involving multiple articles per teaser.

### 5.1 Automatically Identifying Teasers and Matching them to Articles

We describe how we automate the collection of front-page teasers and the process of matching them to the full articles they reference. We examine several methods suited to varying degrees of resource availability.

**Identifying Teasers.** Front-page teasers often include key phrases that indicate that the full story appears elsewhere in the issue. Thus, to identify them automatically we use a rule-based approach that searches for terms specific to each newspaper title. To ensure we include only teasers that function as summaries, we collect additional phrases that signal continuation and use them to filter out non-summary content.

**Matching between front-page teasers and corresponding articles.** Automatically determining which article supports a teaser is challenging. While teasers often point to the relevant page (or pages), the specific article or articles on that page must still be identified.

We develop several methods to perform the matching automatically, spanning from those requiring minimal resources, suitable for low-resource settings, to those utilizing LLMs without requiring any human annotations. We test the proposed methods on Hebrew, which poses unique challenges for LLMs as a morphologically rich language, often resulting in ambiguous structures (Tsarfaty et al., 2020).

**TF-IDF (low resource)**: We follow Ghalandari et al. (2020) that used Term Frequency-Inverse Document Frequency (TF-IDF) to pair news articles with summaries. We generate vector representations for all articles and teasers by computing TF-IDF across the entire corpus, capturing both term frequency and rarity. We then calculate cosine similarity between each teaser and all potential article matches – i.e., the articles on the corresponding page. We define a similarity threshold, based on manually annotated data (50 samples), above which articles are linked to the teaser.

**Finetuned Sentence Transformer (medium resource)**: We use a SentenceTransformer model (Reimers and Gurevych, 2019) to compute similarity scores between teasers and articles by measuring the cosine similarity between their embeddings. A threshold is then applied to determine matches.

| Error Type | Percentage | Category |
|---|---|---|
| Segmentation Error | 66% | False negative |
| Length | 20% | False positive |
| OCR Noise | 13% | False negative |

Table 4: Error analysis of teaser identification using a rule-based approach.

| Method | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|
| TF-IDF | 86 | 93 | 75 | 83 |
| Sentence Transformer | 81 | 82 | 77 | 80 |
| Zero-shot (Llama-3.3 70B) | **90** | **95** | **83** | **88** |

Table 5: Evaluation of automatic teaser-to-article matching. We report accuracy (Acc), precision (Prec), recall (Rec) and F1 score on the Hebrew dataset.

**Zero shot classification (medium to high resource)**: We prompt an LLM to determine for each teaser-article pair whether the teaser summarizes the text. The instruction was given in English, and the text in the original language.

We apply all these methods to a Hebrew newspaper collection. In the next section, we report their quality assessment.

### 5.2 Dataset Quality and Error Analyses

We manually evaluate the automatic process described above finding that it provides a practical option for low-resource settings. All annotations are carried out by native speakers of Hebrew.

**Identifying teasers using a rule-based approach yields strong performance.** After collecting front-page teasers using key phrases, we aim to assess how well this method identifies texts that function as actual summaries. We manually tag 102 texts appearing on front pages to determine whether they could be considered summaries. We then compare these annotations to the results of our rule-based method, which classifies a text as a teaser if it contains a key phrase from a predefined, language-specific, list. We analyze the types of errors in Table 4.

We achieve precision of 95% and recall of 85%. The error analysis reveals that 80% of the errors were false negatives (i.e., teasers that were not identified as summaries), which affects the dataset's scale and coverage but not its quality. These errors are primarily caused by segmentation errors, specifically cases where the front-page teasers text is unavailable to the automatic process because it is marked as an image caption but human annotators can still identify it. Another source of error is OCR noise, where key phrases appear in the front-page teasers but are too distorted to be correctly detected. 20% of the errors are false positives (i.e., texts which were incorrectly marked as teasers), in all of which headers are incorrectly marked as teasers because they follow the teaser format, using the same key word to point to a different page.

**In matching front-page teasers to articles, the zero-shot approach performs best, while TF-IDF presents a viable alternative.** To evaluate the performance of the three proposed methods for matching teasers to summaries – TF-IDF, Sentence Transformers and zero-shot – we manually annotate 50 teasers along with all articles on the pages to which they point. This results in 325 teaser-article pairs, with an average of 6.5 article candidates per teaser. Each pair is annotated to determine whether the teaser can be considered a summary of the article. The annotators overlapped on 110 pairs, yielding a Cohen's kappa (Cohen, 1960) of 0.88, indicating very strong agreement. We then compare our annotations to the classifications produced by each method, with the results presented in Table 5.

The best performance is observed in the zero-shot setting, using instruct tuned Llama-3.3 70B (AI@Meta, 2024). TF-IDF also performs well, outperforming the sentence transformer on most metrics, while requiring minimal computation.

To better understand the capabilities of the low-resource method, we provide an error analysis of the TF-IDF approach in Table 9. The most common error occurs when the article is related to the teaser but not directly supported by it, for example, when both refer to the same event, but the teaser does not explicitly mention it. Other errors occur on noisy samples affected by OCR issues, which a language model is able to handle more robustly, and when the teaser text is short.

### 5.3 HEBTEASESUM: A Real-World Hebrew Summarization Corpus

Our Hebrew dataset, collected using the above method, features a rich variety of summarization shapes and is naturally multi-document.

We apply our method to the "Hadashot" newspaper title, using issues published between 1984 and 1993, available via the National Library of Israel. From a single title, we extract 7,774 samples, of which 3,869 are multi-document (i.e., the summary refers to more than one article). To the best of our knowledge, this is the first available dataset in Hebrew to support multi-document summarization. In Table 6, we compare our data to the available datasets in Hebrew. A detailed summarization datacard and data statistics are provided in Tables 7 and 8. As the human annotation found length serves as an indicator of quality, we also report the number of teasers per length category.

| Dataset | #Samples | Shape | Multi-doc |
|---|---|---|---|
| MassiveSumm | 102,961 | Paragraph | x |
| HeSum | 10,000 | Paragraph | x |
| Ours | 7,774 | Diverse | **3,869** |

Table 6: Comparing our dataset to the available Hebrew summarization datasets: Hesum (Mondshine et al., 2024) and MassiveSumm (Varab and Schluter, 2021).

The contributions of this dataset are two-fold. First, it serves as an example that collecting summarization corpora from digitized newspapers is feasible, with relatively few linguistic resources. Second, the resulting dataset can be used to advance the state of the art in Hebrew NLP. Large-scale data collection remains important for Hebrew, as it enables robust evaluation and supports fine-tuning, for example, to improve morphological understanding (Mondshine et al., 2024). Tables 10 and 11 present benchmark performance on the dataset and the improvements obtained through fine-tuning.

We make the dataset available using our code. To account for OCR errors, we provide a corrected version of the data, achieved by performing prompt-base OCR post correction (Thomas et al., 2024).

## 6 Related work

Many summarization datasets rely on the news domain (Dahan and Stanovsky, 2025) but they are based on web content. Common methods for constructing such datasets include using the article title as a summary (Narayan et al., 2018), extracting the first bolded sentence (Cheng and Lapata, 2016; Hermann et al., 2015), or leveraging social media metadata (Grusky et al., 2018). A few works have focused on identifying summaries and matching them to full articles, which requires an automatic matching process. For example, Ghalandari et al. (2020) linked articles from Wikipedia's Current Events Portal to news articles covering the events.

| Summarization Data Card |
|---|
| <u>**Sample information:**</u> |
| **Languages:** |
| *Hebrew* |
| **Summary Shape:** |
| *Diverse:* Paragraph: 4,277; One-Sentence:2,900; Highlights:597 |
| **Summary Distribution by Length:** |
| 0-25: 2031 |
| 25-50: 1160 |
| 50-100: 2201 |
| >100: 2382 |
| **Domain:** |
| *News* |
| **Size:** |
| *7,774* |
| <u>**Annotation information:**</u> |
| **Annotation efforts:** |
| *Automatic* |
| **Source of supervision:** |
| *Natural* (summaries created organically) |
| **Brief description of the summaries' source:** |
| *Newspapers front page teasers* |
| <u>**Data quality assessment:**</u> |
| **Abstraction level:** |
| *1-gram ratio*: 0.58 |
| *2-gram ratio:* 0.82 |
| *3-gram ratio:* 0.89 |
| *4-gram ratio:* 0.92 |
| **Compression rate:** 0.84 |
| **Human evaluation:** *by domain experts* |
| <u>**Availability details:**</u> |
| **How is the data made accessible:** |
| *URL-based Reconstruction* |
| **Copyrights information:** |
| *License* |

Table 7: Summarization data card.

In contrast to web-based news, newspapers have received limited attention in NLP. However, several works have explored named entity recognition on historical newspapers (Boros et al., 2020; Ehrmann et al., 2020), while others have leveraged NLP methods to analyze historical texts (Borenstein et al., 2023; Candela and Carrasco, 2022).

## 7 Conclusion

We presented a simple method for collecting high-quality summarization data from printed newspapers, demonstrating its suitability for languages with varying levels of resources. The resulting data is abstractive, supports multi-document summarization, and enables evaluation of model capabilities. Using this method, we created a new summarization dataset for Hebrew, HEBTEASESUM, which we release using our code.

| Statistic | Single-Doc | Multi-Doc |
|---|---|---|
| # Teasers | 3,905 | 3,869 |
| Avg. text length (words) | 405 | 312 |
| Avg. summary length (words) | 56 | 102 |

Table 8: Dataset statistics for single and multi-document data. The average number of articles in cluster is 3.4

## Limitations

Although this work supports summarization data collection across diverse languages, we wish to acknowledge several limitations. First, while the method benefits languages with limited online presence, it is less applicable to languages that fell out of use before the 20th century, as front-page teasers only became widespread during that period. Additionally, the quality of the collected data is constrained by OCR errors, which may affect downstream results. For example, highly noisy text can hinder the effectiveness of methods like TF-IDF and may distort evaluation metrics such as novel n-gram overlap.

## Ethical Considerations

Digitized newspapers are provided by national libraries for research purposes, but their content may still be subject to copyright restrictions. To respect these restrictions, we do not publish the collected data online. However, the Hebrew dataset can be reconstructed using the library's API and terms of service, and the multilingual data can be recreated using the list of titles and publication dates provided in the appendix.

## Acknowledgments

We wholeheartedly thank Eyal Miller and the National Library of Israel for granting us access to the archives. We also thank Eliya Habba for the helpful discussions and the anonymous reviewers for their valuable feedback. This work was partially supported by research grant no. 7256 from the Israeli Ministry of Science and Technology.

## References


AI@Meta. 2024. Llama 3 model card.

Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. Weighted krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *EACL 2014*, pages 10–p.

Melodee H Beals, Emily Bell, Ryan Cordell, Paul Fyfe, Isabel Galina Russell, Tessa Hauswedell, Clemens Neudecker, Julianne Nyhan, Mila Oiva, Sebastian Padó, et al. 2020. The atlas of digitised newspapers and metadata: Reports from oceanic exchanges.

Savita Bhat and Vasudeva Varma. 2023. Large language models as annotators: A preliminary evaluation for annotating low-resource language content. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 100–107.

Nadav Borenstein, Karolina Stańczak, Thea Rolskov, Natália da Silva Perez, Natacha Klein Käfer, and Isabelle Augenstein. 2023. Measuring intersectional biases in historical documents. *arXiv preprint arXiv:2305.12376*.

Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 431–441, Online. Association for Computational Linguistics.

Gustavo Candela and Rafael C Carrasco. 2022. Discovering emerging topics in textual corpora of galleries, libraries, archives, and museums institutions. *Journal of the Association for Information Science and Technology*, 73(6):820–833.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Noam Dahan and Gabriel Stanovsky. 2025. The state and fate of summarization datasets: A survey. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7259–7278, Albuquerque, New Mexico. Association for Computational Linguistics.

Melissa Dell, Jacob Carlson, Tom Bryan, Emily Silcock, Abhishek Arora, Zejiang Shen, Luca D'Amico-Wong, Quan Le, Pablo Querubin, and Leander Heldring. 2023. American stories: A large-scale structured text dataset of historical us newspapers. *Advances in Neural Information Processing Systems*, 36:80744–80772.

Daniel Deutsch and Dan Roth. 2021. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309.

Maud Ehrmann, Matteo Romanello, Stefan Bircher, and Simon Clematide. 2020. Introducing the clef 2020 hipe shared task: Named entity recognition and linking on historical newspapers. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 524–532. Springer.

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.

Jessica Forde, Ruochen Zhang, Lintang Sutawika, Alham Aji, Samuel Cahyawijaya, Genta Indra Winata, Minghao Wu, Carsten Eickhoff, Stella Biderman, and Ellie Pavlick. 2024. Re-evaluating evaluation for multilingual summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19476–19493.

Shengxiang Gao, Fang Nan, Yongbing Zhang, Yuxin Huang, Kaiwen Tan, and Zhengtao Yu. 2025. A mixed-language multi-document news summarization dataset and a graphs-based extract-generate model. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9255–9265, Albuquerque, New Mexico. Association for Computational Linguistics.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. *Preprint*, arXiv:2005.10070.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. Metal: Towards multilingual meta-evaluation. *arXiv preprint arXiv:2404.01667*.

Mohamed Haneefa and PT Jiji. 2019. Contents and interactivity of national library websites. *DESIDOC Journal of Library & Information Technology*, 39(3):131.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Kimmo Kettunen, Mika Koistinen, and Jukka Kervinen. 2020. Ground truth ocr sample data of finnish historical newspapers and journals in data improvement validation of a re-ocring process. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 30(1):1–20.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Haoyuan Li, Yusen Zhang, Rui Zhang, and Snigdha Chaturvedi. 2024. Coverage-based fairness in multi-document summarization. *arXiv preprint arXiv:2412.08795*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yixin Liu, Alexander R Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, et al. 2022. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *arXiv preprint arXiv:2212.07981*.

Itai Mondshine, Tzuf Paz-Argaman, Asaf Achi Mordechai, and Reut Tsarfaty. 2024. Hesum: a novel dataset for abstractive text summarization in hebrew. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 26–36.

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

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *Preprint*, arXiv:2309.09558.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Chantal Shaib, Joe Barrow, Alexa F Siu, Byron C Wallace, and Ani Nenkova. 2024. How much annotation is needed to compare summarization models? *arXiv preprint arXiv:2402.18756*.

Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. 2021. How well do you know your summarization datasets? *arXiv preprint arXiv:2106.11388*.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging llms for post-ocr correction of historical newspapers. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 116–121.

Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From spmrl to nmrl: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (mrls)? *arXiv preprint arXiv:2005.01330*.

Sandra H Utt and Steve Pasternack. 1989. How they look: An updated study of american newspaper front pages. *Journalism Quarterly*, 66(3):621–627.

Daniel Varab and Natalie Schluter. 2021. Massivesumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023b. Zero-shot cross-lingual summarization via large language models. *arXiv preprint arXiv:2302.14229*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can nlp help revitalize endangered languages? a case study and roadmap for the cherokee language. *arXiv preprint arXiv:2204.11909*.

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

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

# A Newspaper Use

We provide links to the original newspapers from which the front-page teasers examples in Figure 2 were taken and their translations, along with the issue dates used to collect front-page teasers for Section 3. We also provide a list of key phrases used to identify teasers in various languages.

## A.1 Picture Origin:

1. Rana Blad (Norway)
2. Stampa Sera (Italy)
3. Fréttablaðið (Iceland)

## A.2 Translations:

Translations of the teasers shown in Figure 2 are provided to illustrate the data quality, generated using GPT-4o.

1. Today the National Ski Championship opens, and the first day may belong to the recruits. Elin Nilsen from B&Y IL, and Inger Lise Hegge from Henning, are both recruits who earlier in the season have challenged the national team skiers. The women will race 10 kilometers today, while the men will do 30 kilometers. In addition, there is the combined ski jumping and team jumping events.
2. Hopes of forming a national government in Macedonia diminished significantly yesterday, as an Albanian political party demands that attacks on armed Albanian residents in Kosovo be halted before talks begin.
3. Alessandra Casella, 27 years old, is the co-star of Fredi, currently showing at the Alfieri Theater in Turin. After her experience at the Actor's Studio and lots of theater, she would like to work more in cinema, but as she told us (interview on page 27), "in Italy, times are tough — directors prefer fashion models."

## A.3 Dates for collection:

For the data collection in Section 3, we randomly select a starting date and review all subsequent issues of the title, following the library's order, which was sometimes not chronological. We present dates here.

1. **Rana Blad (Norway)**: 01.02.1990-10.02.1990
2. **Kathimerini (Greece)**: 01.12.2010, 02.11.2010, 01.10.2010, 01.09.2010, 17.08.2010, 01.07.2010, 01.05.2010, 01.06.2010
3. **Stampa Sera (Italy)**: 09.10.1991-18.10.1991
4. **Eesti Päevaleht (Estonia)**: 01.01.2018, 31.01.2018, 02.02.2018, 03.03.2018, 02.03.2018, 01.04.2018, 02.04.2018, 01.05.2018
5. **Fréttablaðið (Iceland)**: 23.04.2001, 24.04.2001, 26.04.2001, 30.04.2001, 02.05.2001, 03.05.2001, 04.05.2001, 07.05.2001, 08.05.2001, 09.05.2001, 10.05.2001, 11.05.2001
6. **Hadashot (Israel)** 01.01.1992 - 08.01.1992
7. **Dziennik Polski (Poland)**: 01.01.2002, 03.01.2002, 05.01.2002, 07.01.2002, 08.01.2002, 09.01.2002, 10.01.2002

## A.4 Key phrases

For the manual collection we used the following key phrases to identify teasers:

- **Norwegian**:Side. ("Page")
- **Icelandic**:Bls. (abbreviation for "Page")
- **Estonian**:LK. (abbreviation for "Page")
- **Greek**: Sel, written in Greek script. (abbreviation for "Page")
- **Hebrew**: Am, written in Hebrew script. (abbreviation for "Page")
- **Italian**: Pag. (abbreviation for "Page")
- **Polish**: Szczegóły and str. (Details and abbreviation for "Page", respectively)

| Error Type | Percentage |
|---|---|
| Related articles | 39% |
| OCR noise | 32% |
| Others | 24% |
| Length | 4% |

Table 9: Error analysis of the TF-IDF approach to match teasers to articles. Out of 325 pairs we find 41 errors.

## B Teasers Examples

We present here translated examples of teasers from our dataset.

**Under 25 words:** "Children Destroyed 2.6 Million Shekels in Search of Smurf Stickers."

**between 25 and 50 words:** "Thousands Spent the Holiday Away from Home: 25,000 Travelers Flocked to the Beaches of Eilat. The hotels, hostels, and especially the beaches were filled to capacity. The beaches of the Sea of Galilee were also crowded, particularly Tzemach Beach, where a rock festival was taking place. About ten thousand people passed through the Taba terminal and traveled to Sinai."

**between 50 and 100 words:** "Driver Who Ran Over Child During 'Road Roulette' Acquitted. Boris Eligolashvili of Ramla, who was accused of causing the death of 11-year-old Baruch Oren from Haifa, was acquitted yesterday in the Haifa Traffic Magistrate's Court. The judge ruled that the accident was not caused by his negligence. Baruch Oren was killed on March 17, 1989, on the Haifa–Tel Aviv highway, near Kiryat Shprintzak in Haifa. Oren's friends said that he was waiting in the median strip of the road, and when Eligolashvili's car was about a hundred meters away, he jumped onto the road and stood in front of it with his arms folded."

**over 100 words:** "120,000 Vaccinated Against Polio in Hadera Governorate as of This Morning; Two Additional Cases Suspected. The Ministry of Health is launching an unprecedented campaign this morning, during which 120,000 people in Hadera Governorate will be vaccinated against polio. These are individuals aged 0 to 34 years. The vaccine, which protects recipients from the virus, will be administered orally using the Sabin vaccine. Pregnant women and infants who have never been vaccinated will receive an injection of the Salk vaccine. Today, about a thousand infants and all pregnant women in Hadera Governorate will be vaccinated. The day after tomorrow and on Friday, the Sabin vaccine will be administered to about 50,000 kindergarten-age children and students in the education system. Starting next Sunday, the adult population and children below kindergarten age will be vaccinated. The Ministry of Health announced yesterday two new suspected cases of polio: a 27-year-old resident of Hadera and a young man, about 24 years old, from Zichron Yaakov. The condition of the nine-month-old baby from Kiryat Gat, who was brought to Barzilai Hospital three weeks ago with suspected polio, has improved."

## C Human annotation

We follow Fabbri et al. (2021) annotation guidelines.

"In this task you will evaluate the quality of summaries written for a news article. To correctly solve this task, follow these steps:

1. Carefully read the news article(s), be aware of the information it contains.
2. Read the proposed summary
3. Rate each summary on a scale from 1 (worst) to 5 (best) by its relevance, consistency, fluency, and coherence.

Definitions:

Relevance: The rating measures how well the summary captures the key points of the article. Consider whether all and only the important aspects are contained in the summary. When multiple articles are presented, the score should account for the most important information across all sources.

Consistency: The rating measures whether the facts in the summary are consistent with the facts in the original article. Consider whether the summary does reproduce all facts accurately and does not make up untrue information.

Fluency: This rating measures the quality of individual sentences, are they well-written and grammatically correct. Consider the quality of individual sentences.

Coherence: The rating measures the quality of all sentences collectively, to the fit together and sound naturally. Consider the quality of the summary as a whole. "

## D Prompts

We provide the prompts used in our work.

| | Model | ROUGE | | | BERTscore | LLM-as-a-Judge (1-5 scale) ↑ | | |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | L | | Coherence | Consistency | Coverage |
| Single-doc | Mixtral | 0.11 | 0.02 | 0.10 | 0.68 | 2.79 | 2.81 | 2.07 |
| | Llama | **0.15** | **0.05** | **0.14** | 0.69 | 4.00 | 4.29 | 2.49 |
| | DeepSeek-R1 | 0.13 | 0.03 | 0.12 | **0.70** | **4.50** | **4.74** | **2.65** |
| Multi-doc | Mixtral | 0.12 | 0.02 | 0.11 | 0.67 | 2.67 | 2.77 | 1.94 |
| | Llama | **0.16** | **0.04** | **0.14** | **0.69** | 3.88 | 4.19 | 2.27 |
| | DeepSeek-R1 | 0.12 | 0.02 | 0.11 | 0.68 | **4.41** | **4.73** | **2.29** |

Table 10: Model performance on HEBTEASESUM.

| model | ROUGE | | | BERTScore |
|---|---|---|---|---|
| | 1 | 2 | L | |
| mLongT5 | 0.08 | 0.02 | 0.07 | 0.57 |
| mLongT5 (FT) | 0.16 | 0.07 | 0.16 | 0.68 |

Table 11: Performance gains on mLongT5-base (Xue et al., 2021) following 10 fine-tuning epochs.

**Prompt use for summarization task:** We follow Wang et al. (2023b). "Please summarize the following text in [LANGUAGE]:

Text: [TEXT]

**Prompts use for LLM-as-a-judge:**

- **Coherence:** We follow Wang et al. (2023a). "Score the following news summarization given the corresponding news with respect to coherence with one to five stars, where one star means “incoherence” and five stars means “perfect coherence”. Note that coherence measures the quality of all sentences collectively, to the fit together and sound naturally. Consider the quality of the summary as a whole."

- **Consistency:** We follow Wang et al. (2023a). "Score the following news summarization given the corresponding news with respect to consistency with one to five stars, where one star means “inconsistency” and five stars means “perfect consistency”. Note that consistency measures whether the facts in the summary are consistent with the facts in the original article. Consider whether the summary does reproduce all facts accurately and does not make up untrue information."

- **Coverage:** We follow Liu et al. (2022). "You will receive a reference summary and a candidate summary. Your task is to compare these two summaries and assess the extent to which the candidate summary covers the information presented in the reference summary.

  Please indicate your agreement with the following statement: “All of the information in the reference summary can be found in the candidate summary.”

  Use the following 5-point scale when determining your response:

  1. Strongly Disagree
  2. Disagree
  3. Neither Agree nor Disagree
  4. Agree
  5. Strongly Agree

  Reference Summary:referece

  Candidate Summary: generated

  Evaluation Form (scores ONLY): - Agreement (1-5):"

**Prompt use for matching teasers to articles:** "Given the following text and summary, answer with ’Yes’ if the text relates to the summary, and ’No’ if it does not. Do not provide explanations. Only output ’Yes’ or ’No’."