Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Filling the Blanks: Context-Driven Detection and Correction of Cherry-Picking in News Reporting

Israa Jaradat, Haiqi Zhang, Chengkai Li
University of Texas at Arlington
Arlington, TX, USA
{israa.jaradat,haiqi.zhang}@mavs.uta.edu, cli@uta.edu

Abstract

Cherry-picking involves suppressing (censoring) or distorting evidence that supports the counter argument. Cherry-picking facts in news reports by mainstream media distorts public perception, undermines trust, and fuels misinformation by presenting a biased or incomplete narrative. Manually identifying suppressed statements in news stories is challenging and time consuming. In this study, we introduce a novel importance-based approach to automatically spotting and correcting cherry-picking by identifying then substituting missing important statements in a target news story with the help of contextual information from other news sources with different biases. Additionally, we showcase the flexibility of our approach by utilizing different methods to estimate a statement's importance including fine-tuned embedding models, zero and few-shot generative models, in addition to unsupervised methods. Furthermore, this research introduces a novel dataset specifically designed for training and evaluating cherry-picking detection methods. Our best performing method achieves an F-1 score of above 90% in estimating a statement's importance. Moreover, results show the effectiveness of the proposed approach in correcting cherry-picking and bringing the biased narrative closer to its neutral alternative by nearly 12%. Finally, through thorough experimentation, we provide answers to a set of important research questions related to cherry-picking detection and correction.

Introduction

The erosion of citizens' trust in the media has reached a critical stage, posing significant harm to the foundations of democracy (Dan et al. 2021; Druckman and Parkin 2005; Chiang and Knight 2011). Research shows evidence of a decline in media trust among citizens (Strömbäck et al. 2020). Media losing trust can be attributed to the perception of political bias in media reporting (Lee 2010) and consumers' exposure to disinformation (Lee, Gil de Zúñiga, and Munger 2023; Hameleers, Brosius, and de Vreese 2022).

Cherry-picking, a widely used but often imperceptible practice in news reporting, is a device that aids bias and disinformation (Vincent 2019; Paul and Elder 2019). It is defined in literature as a logical fallacy that involves deliberately censoring or slanting crucial facts that support the opposing viewpoint (Best 2004). Its common aliases in literature include *one-sideness* (Paul and Elder 2008a), *argument by half-truth* (Sproule 2001), *case-making* (Lee 1945), and *censorship* (Gläßel and Paula 2020).

Cherry-picking is prevalent across various domains, such as media narratives, science and politics (Best 2004). In news reporting, cherry-picking is observed across scales as large as in selecting which events to cover, and as subtle as in choosing specific words in a news report. Furthermore, cherry-picking manifests in various types of data. For example, for facts staffed with numbers and stats, specific segments of a trend-line (Asudeh et al. 2020) or particular statistics are left out to bolster a particular claim (Wu et al. 2017). Empirical studies in media and communication research have identified several modalities of cherry-picking, including the exclusion of contradictory statistics (Best 2004), the deliberate neglect of relevant historical or socio-political context (Paul and Elder 2008b), and the preferential citation of experts whose views support a predetermined angle (Rodrigo-Ginés, Carrillo-de Albornoz, and Plaza 2024). Moreover, other studies identify another form of cherry-picking involving over-emphasizing (i.e., exaggeration) of facts or a certain perspective (Paul and Elder 2008b).

This paper aims at studying the automation of detecting and correcting cherry-picking of statements within news reports, particularly when certain statements or facts are suppressed or censored from a news story. We assume that a sentence in a news report carries a single fact since it is a common practice in news reporting to represent a singular idea or fact in one sentence (Mencher and Shilton 1997). Media outlets, including prominent and highly regarded ones, exhibit biases in various topics within domains such as politics and science (Baron 2006) by selectively censoring **important** (i.e, key) statements, particularly those that challenge their biases (Paul and Elder 2008a; Vincent 2019; Brown 1963). This soft-censorship (Podesta 2009) is used to manipulate the public opinion on pivotal subjects, such as climate change, vaccination, or elections, by misleading readers to perceive partial truth as holistic (Fleming 1995). Specifically, this study strives to answer the following series of research questions, by proposing and evaluating an importance-based novel approach for automating cherry-picking detection and correction: **(Q.1)** Can important statements censored from a given news report be iden-

tified using context derived from other sources' news stories? **(Q.2)** What are the best means of leveraging this context to automatically spot important censored statements? **(Q.3)** How much context is effective? **(Q.4)** Is context derived from sources with counter-biases (i.e., with the absence of a neutral narrative) enough to effectively spot censored statements? **(Q.5)** Can we mitigate cherry-picking in a biased news story by integrating missing statements from the counter perspective? **(Q.6)** What levels of a statement's importance granularity can be effectively estimated given the right context? **(Q.7)** Are cherry-picking and general bias correlated in news sources?

To answer the above questions, we introduce a novel importance-based approach for identifying and correcting censored statements in news stories by leveraging context derived from other narratives. Our approach is flexible enough to utilize a variety of language models, including fine-tuned embedding models and few and zero-shot prompting of generative models, in addition to unsupervised models. These models are used to assess statements' importance based on context derived from other news stories to identify censored statements. The models are trained and evaluated on the novel *Cherry* dataset we curated for this study. Additionally, our approach uses the spotted censored statements to correct cherry-picking by providing an alternative more neutral news story. The proposed approach shows promising results in spotting and correcting cherry-picking, given external context.

While computational methods have advanced the detection of various forms of media bias, including ideological slant and framing (Baly et al. 2020; Spinde et al. 2021b), relatively less attention has been given to the automated detection of cherry-picking. By framing cherry-picking as a problem of biased omission, our work extends the landscape of computational media bias detection. Furthermore, we harness large language models (LLMs) not only to estimate the importance of omitted statements, but also to generate context-informed corrections.

To the best of our knowledge, this is the first study that addresses the detection and correction of cherry-picking in text using computational approach. The key contributions of this work are as follows: (a) The formulation and modeling of the problem from a computational perspective. (b) The novel end-to-end importance-based cherry-picking detection and correction approach that infuses contextual information with statements through different language modeling techniques, and provides a more balanced alternative to a given biased news story. (c) A new cherry-picking annotated dataset which can prove useful in not only detection and correction but also other lines of research related to cherry-picking. (d) Thorough experimentation to answer the previous research questions which involve comparing between different models and determine the most effective context to consider when assessing cherry-picking. All artifacts of this work are released under the GNU General Public License v3.0 at our anonymous Github repository https://github.com/cherry-pic/Cherry.

Related Work

Media Bias Granularity. Research on media bias can be classified into three levels based on text granularity: word-level, sentence-level, and document-level (Chen et al. 2020). Distinguishing between these levels is not always straightforward, as they often intersect and influence each other. For instance, word-level bias is dependent on sentence or document context, and similarly sentence-level bias is contingent on the broader article context. **1) Word-level Bias.** The choice of words in news stories can influence the audience's opinions towards a particular event (Hamborg 2020). Word-level bias encompasses two primary categories: framing bias and epistemological bias (Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013). The connotations associated with the same word can vary by the specific context in which they are employed (Spinde et al. 2021b). **2) Sentence-level Bias.** Numerous studies focus on detecting bias at sentence level, considering it as the fundamental level that can be combined with document-level bias (Spinde et al. 2021a). Using lexicon of bias words, Hube and Fetahu (2018) introduced a method for identifying biased statements from Wikipedia. In Lei et al. (2022), biased sentences were identified in news articles, serving to explain the overall bias present throughout articles. Focusing on media's sentiment toward target entities, Fan et al. (2019) examined sentence-level bias along with article context. **3) Document-level Bias.** Baly et al. (2020) constructed a dataset of news articles annotated with left, center, and right leanings, then used machine learning models to predict articles' political ideology.

Cherry-picking. While existing media bias frameworks such as framing bias and lexical bias operate primarily at the level of language, cherry-picking represents a more covert form of bias—bias by omission or over-emphasis. Specifically, cherry-picking does not modify or frame existing content but instead selectively removes entire statements or facts, often those that contradict a desired narrative. In contrast to methods that highlight what is said, our approach identifies what is not said—providing a valuable addition to the suite of tools for analyzing media bias. Prior research on media bias have left a notable gap in exploring media bias related to cherry-picking in news reporting. Although some studies have examined the impact of cherry-picking in computational fact-checking (Wu et al. 2017; Lin et al. 2021) and the detection of cherry-picking in trendlines (Asudeh et al. 2020), there is no specific investigation into end-to-end cherry-picking detection in news reporting.

LLM-based Media Bias Detection. Recent advancements in LLMs have catalyzed progress in automating media bias detection across multiple dimensions, including framing (Lin et al. 2024; Wang et al. 2025), lexical bias (Horych et al. 2024), and other biases (Shah, Shah, and Attar 2024). Inspired by these studies, our work applies LLMs to identify cherry-picking—an omission-based bias—through context-aware importance estimation and corrective integration.

Cherry-picking Statements in News Stories: Problem Formulation and Modeling

Media outlets cherry-pick in news reporting by censoring important relevant statements, especially those against their biases (Paul and Elder 2008a; Vincent 2019; Brown 1963). To offer a more illustrative portrayal of the problem, refer to a detailed juxtaposition of three news reports from sources with different biases covering the same event in Table 6 in the Appendix.

The example in Table 6 demonstrates two important insights that guide our approach. *First*, censored statements can be potentially discerned by their importance to the event being covered. Therefore, we exploit statement importance in detecting censored statements. We opt for assessing cherry-picking on the statement (i.e., sentence) level since it is a common practice in news reporting to represent a singular idea or fact in one sentence (Mencher and Shilton 1997). Note that stories often face constraints on space due to factors such as reader's attention span, which results in prioritizing the most important statements to appear in a news story. *Second*, the importance of a statement is relevant, contextualized by whether it is discussed or not in other stories about the same event—simply put, together all stories depict a comprehensive picture. For a comprehensive and representative context, it is crucial to incorporate narratives from other sources (Lee and Lee 1939).

Drawing from the discussion above, the problem of identifying censored statements is formulated as follows. Consider an event e=$\{d_1, ..., d_n\}$ comprising various narratives (i.e., documents). Each document d_i=$\{s_1, ..., s_m\}$ contains a collection of statements (i.e., sentences). The statements from all documents in e collectively form the universal set of statements S_e=$\{d_1 \cup d_2 ... \cup d_n\}$ for the entire event. Our goal is to determine the set of important statements that are missing (censored) from each document, i.e., c_i=$I_e - d_i$ in which $I_e \subset S_e$ represents the important statements regarding e, among all statements. The problem then reduces to the task of finding I_e based on the event's context, which can be approached as a classification task. This reduction enables our approach to adapt to a variety of models to find I_e. In this study we use news reports from distinctively biased sources covering event e as its context d. This is based on the need for the counter-perspective to identify cherry-picking (Lee and Lee 1939). To determine source bias, we utilized categorization of media sources bias based on unanimous agreement from three sources—Media Bias Fact Check (MBFC),[1] AllSides,[2] and Ad Fontes Media.[3] These sources are widely utilized in media bias analysis, providing researchers with tools and ratings to evaluate the political bias and credibility of news sources. The details of how they assess bias in news outlets can be found in Appendix.

[1]https://mediabiasfactcheck.com/

[2]https://www.allsides.com/

[3]https://adfontesmedia.com/

Dataset

To train and evaluate our models for gauging statement importance given event context, we curated a novel cherry-picking detection dataset *Cherry*. Our dataset is composed of 3,346 examples in total. Every example contains a statement $s \in S_e$ (i.e., a single sentence from a news report of event e), an event context $d \in e$ which contains a news report covering the event collected from a different source, and an importance label Y that indicates whether the statement is important to the event or not.

Data Collection

To generate the examples in the dataset, we compiled a list of 41 noteworthy news sources, such as CNN.com, Reuters.com, and FoxNews.com, selected for their high publication frequency and distinct political biases as determined by the ratings from the aforementioned three websites. Subsequently, we employed GDELT's API (Leetaru and Schrodt 2013) to gather all news reports except for opinion articles and editorials from the chosen sources. The collected news reports cover the time span between December 2019 and January 2021.

Next, we used DBSCAN (Ester et al. 1996) to cluster the collected news reports into events based on the cosine similarity calculated between the vectors of the reports, with each report being vectorized through the concatenation of BERT (Devlin et al. 2019) and TF-IDF representations derived from its headline and initial paragraph. We set DBSCAN's parameters to 0.04 for the radius of neighborhood around data points (ε) and 2 for the minimum cluster size. We tuned DBSCAN parameters by manually inspecting the clustering performance with different combination of these two parameters on a sample of events that were published in the same day.

After a manual inspection of the generated clusters, we curated a subset of 82 clusters, each corresponding to a distinct controversial event. We chose these events based on the total number of news reports published from each source and bias category assuming that events with higher number of published news reports are likely talking about a major controversial event. Examples of such events include the January 6th, 2021 U.S. Capitol attack and the alleged foreign intervention in the 2020 U.S. presidential elections. [4]

For each event, we used NLTK (Bird, Klein, and Loper 2009) sentence tokenizer to segment the news reports into statements, and then clustered the statements based on semantic similarity. Similar to how reports were clustered, we used BERT and TF-IDF to represent statements and applied DBSCAN with a value of 0.07 for ε, a minimum cluster size of 2, and cosine similarity of statement vectors as the similarity function. DBSCAN parameter values were tuned based on the best accuracy we could achieve from the different combinations of the two hyper-parameters on a data sample of 508 statements that were labeled with cluster IDs manually for this purpose. The sample used for tuning these

[4]The full lists of the 82 events and the 41 sources are available at https://github.com/cherry-pic/Cherry.

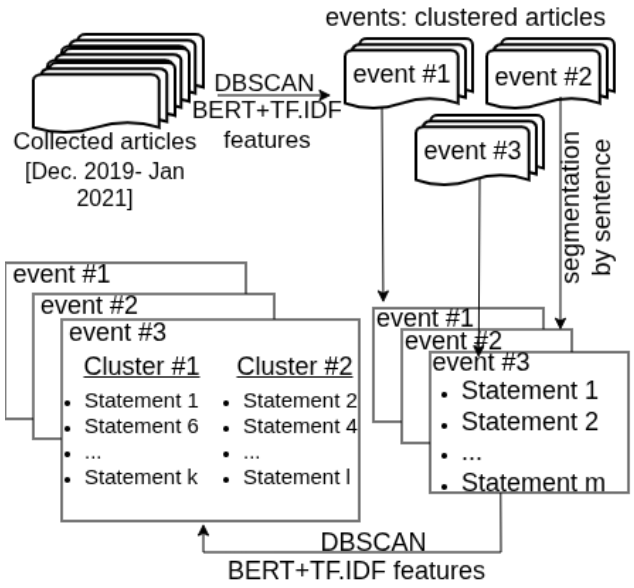

Figure 1: Data collection and preparation pipeline.

two parameters are publicly available on our Github repository. Table 7 in Appendix shows a sample of clustered statements within a single event. Statement clustering facilitates data augmentation on the dataset, where the collected label for a single statement was cast over all statements in the same cluster. As a result, multiple examples were labeled with a single example labeling effort.

The data collection, clustering, and segmentation pipeline is depicted in Figure 1. After clustering the statements, we fed them along with their respective context into our custom data annotation tool.

Data Annotation

We developed a data annotation tool to collect labels for statements that reflect the statements' importance to corresponding events. Refer to Figure 7 in the Appendix for a screenshot of the our annotation tool. The annotators comprise one faculty member and nine Ph.D. students from a university research lab. They are diverse in terms of political direction, cultural background, nationality, gender, and age.

For each event, the annotators are presented with a news report from a least-biased source (e.g., Reuters.com), along with clusters of statements related to the event. The annotators are required to thoroughly read the news reports and gain a comprehensive understanding of the event covered. They are instructed to assign one of the following five labels to the cluster: (1) *very important*, (2) *kind of important*, (3) *not very important*, (4) *the excerpts might be incorrect*, and (5) *I am not sure*. These labels convey the perceived level of importance of the statements in the cluster.

Following the collection of labels from annotators, we filtered out examples with less than 3 annotators or less than 75% agreement ratio. Agreement ratio is the number of majority votes divided by the total number of annotators for an example).

Dataset Statistics

Considering the inherent difficulty in delineating the decision boundary between the different labels described in the previous subsection, we opted to explore various binary and multinomial classification configurations as depicted in Table 1. This involved merging certain labels into a single class or excluding specific labels from the dataset. The class distribution for each of the four classification configurations is provided in Table 1. For each configuration, we split the dataset randomly into 85% for training and 15% for testing.

Each example in the dataset consists of labels, context collected from other sources news stories, and statements. Multiple statements stem from the same event, and consequently the context repeats within the dataset examples. Hence, we split the dataset between testing and training by events instead of stratification to avoid any data leakage (i.e., model's exposure to event contexts from the testing dataset during training).

Conf.	Class 1	Class 2	Class 3
1	{(1)}	{(2),(3),(4),(5)}	-
	2175 (64%)	1232 (36%)	-
2	{(1)}	{(2),(3)}	-
	2175 (65%)	1171 (35%)	-
3	{(1)}	{(2),(3)}	{(4),(5)}
	2175 (64%)	1171 (34%)	61 (2%)
4	{(1)}	{(2)}	{(3)}
	2175 (65%)	667 (20%)	504 (15%)

Table 1: Label combinations and class distribution (number of examples and ratio with regard to the whole dataset) for each of the four classification configurations. The label for each index from (1) to (5) can be found in the previous subsection.

Methodology

Our importance-based approach identifies censored statements c_i in a given news report d_i using context derived from other sources news stories d, as follows:

$$c_i = \{s \in S_e : f(s, d) = 1\} - d_i \quad (1)$$

where f is a classification method which assigns a class of 1 for important statements and 0 for unimportant ones. The pipeline finds c_i by first segmenting all documents in event e to obtain the universal set of statements S_e. Then each statement in S_e is scored for importance using a classification method and the event context document d. Finally, each important statement is verified for its presence in each document (i.e., news report) using semantic similarity or an LLM skill (See Statement existence verification skill in Appendix). If the statement is absent in a document, it is appended to the list of censored statements associated with that document.

After identifying the set of missing important statements c_i, from document d_i we integrate them into the document.

This integration aims to mitigate cherry-picking by providing an alternative document d_i' that is closer to a neutral narrative d_n in the embedding space, such that:

$$cos(\vec{d_i}, \vec{d_n}) < cos(\vec{d_i'}, \vec{d_n}) \quad (2)$$

$$\text{where } d_i' = d_i + c_i \quad (3)$$

The integration of c_i into d_i is achieved through an LLM skill (See Statement integration skill in Appendix for the skill's template) that places each missing statement in is most suitable place within d_i while maintaining the original text (i.e., linguistic and stylistic characteristics).

Classification methods $f(s, d)$

We introduce three different methods that serve the purpose of estimating the importance of a statement with regard to an event, given its context. Our methods emulate the approach taken by human annotators—reading the context about the event and then assessing the statement's importance considering the context.

Fine-tuned Supervised Models To consider context while estimating a statement's importance, we fine-tuned supervised language models using the sequence pair classification task on our dataset as illustrated in Figure 2. A sequence pair consists of a statement and the corresponding context. The context is taken from other news stories within the event collection. If the token limit of a certain language model is exceeded, the sequence is right-truncated. The encoded sequence is represented as a vector obtained from the output of the [CLS] token. This vector is subsequently passed through a linear layer to produce a class probability.

We employed both BERT (Devlin et al. 2019) and Longformer (Beltagy, Peters, and Cohan 2020) as our supervised models. Unlike BERT, Longformer exhibits linear scaling with input sequence length. As a result, it enables the processing of lengthier documents, accommodating approximately 4096 tokens compared to BERT's limit of 512 tokens. This capability aligns well with our requirement for typically long input sequences of context (i.e., full news reports). The primary factor that distinguishes Longformer as a computationally efficient variant from BERT is its novel self-attention mechanism. Longformer applies self-attention over a sliding window or dilated window of tokens, instead of every token in the input sequence. Additionally, to learn task-specific representations, Longformer uses global attention on preselected tokens form the input sequence. These tokens attend to every other token in the sequence, and every token in the sequence attends only to these preselected tokens, which facilitates the learning of task-specific representations (Beltagy, Peters, and Cohan 2020), further enhancing its suitability for our needs.

Large Language Models with Zero-shot or Few-shot Demonstrations The advantage of zero-shot and few-shot demonstration is that it does not require large amount of data or human supervision. In utilizing the generative language model GPT (Brown et al. 2020) to assess a statement's importance, our approach imitates the way human coders annotated the Cherry dataset. We use the same prompt in the

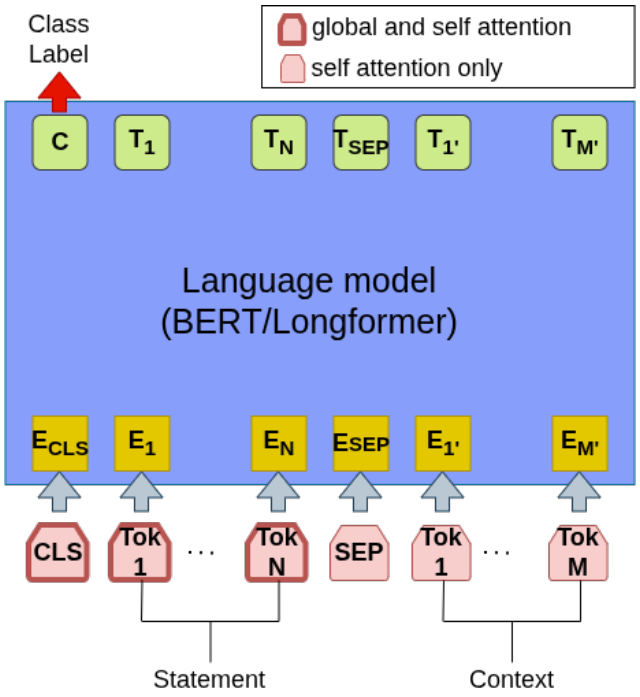

Figure 2: Automatic cherry-picking detection using sequence-pair classification architecture, showing the locations of global attention in the input sequence.

annotation interface (i.e., the annotation instruction) with slight modifications (refer to Appendix for the full template). We first let the model read the context collected from other news sources' coverage for the event. We then prompt the model to answer whether the statement is important to mention in a news story covering the event. Alternatively, a few demonstrations can be embedded within the prompt for the model to learn from, where each demonstration example contains the context document, the statement to assess, the question asking the model if the statement is important considering the context news story, and the answer to the question (i.e., Yes/No). The full prompt template is included in Appendix.

Unsupervised Methods Unsupervised summarization methods are capable of ranking sentences in a given textual corpus without the need for human supervision. These summarizers score then rank sentences in the given text by based on their importance and generate a summary of size s by giving the most s important sentences. These methods can be utilized to find I_e (important statements regarding event e) with less cost and no human supervision effort compared to the other approaches. For this purpose, we can utilize classic summarization algorithms such as LexRank (Erkan and Radev 2004) and TextRank (Mihalcea and Tarau 2004), or even more recent ones that utilize language models such as BERT-based summarizers (Miller 2019) and LLM-based rankers. LexRank is fit by constructing a graph with sentences as node and edges are weighted based on the similarity between sentences. The

centrality of each node is then calculated from a stochastic matrix of these weights, then given a new set of sentences. LexRank extracts a specific length summary by ranking the sentences based on their centrality scores. TextRank builds a graph where sentences are nodes, and edges represent their similarity. By applying a version of the PageRank algorithm, it scores each sentence based on its connections, identifying the most important sentences. In (Miller 2019), BERT is used to vectorize sentences after tokenizing input text. Vectors then are clustered with K-means and the closest sentences to the centroids of clusters are considered in the summary. While LLMs can be used as classifiers as we describe in the previous section (i.e., pointwise ranking), they can be used as rankers as well to extract a summary of size s sentences from the text by comparing sentences with each other. For the purposes of this task, LLMs can be utilized as rankers by tokenizing the text into sentences, then feeding them along with the text to the LLM and prompting it to rank the sentences by their importance with regard to the text.

Experimentation and Results

We designed and implemented a set of experiments, in order to answer the research questions we brought up earlier using our proposed methodology and dataset described in earlier sections. All the code base, experiments and results are available at our Git repository: https://github.com/cherry-pic/Cherry.

(Q.1) Can important statements censored from a given news report be identified using context derived from other sources' news stories? To answer this question, we utilized a BERT-base[5] model as our **baseline**, with a statement alone (i.e., no context) as its input sequence. We also created a variant of the model by forming the input sequence as the statement and its context concatenated and separated by the [SEP] token. Context in this experiment is collected from a neutral source (i.e., Reuters.com). We use macro F-1 and accuracy as models' performance measures in this and subsequent experiments. Our hypothesis is that, just as humans can make more accurate judgments regarding the importance and cherry-picking of a statement when provided with an unbiased context (Lee and Lee 1939), models are anticipated to exhibit similar capacity. The baseline model scored 0.619 and 0.617 in accuracy and F-1, respectively. When the same model was given 100 words of context to attend to when classifying statements, its accuracy and F-1 were significantly improved to 0.846 and 0.870, respectively. Note that language models can still capture some signals pertinent to statements' importance from their content without context (thus better than a random guesser).

(Q.2) What are the best means of leveraging this context to automatically spot important censored statements? and (Q.3) How much context is effective? Given the importance of context as established above, we answer these two questions by comparing the performance of various proposed methods for estimating a statement's importance at different context lengths. Toward this, we trimmed the context in both the training and test sets at different lengths measured in words, and we then fine-tuned the supervised models and evaluated all models. Similar to the experiment above, context in this experiment is collected from a neutral source.

For supervised models, we used the smaller variant BERT-base consisting of 12 transformer layers, 12 attention heads, and 110M parameters, in addition to Longformer-base [6] consisting of 12 transformer layers, 12 attention heads and 149M parameters. We fixed the learning rate at 2e-05, the batch size at 8, and the classification threshold at 0.5. To maintain performance and efficiency, we set global attention in Longformer models at the classification and statement tokens only as shown in figure 2. For details of experimentation on global attention locations refer to Appendix. Additionally, we set the number of epochs for training the supervised models to 5, which required at most five minutes of training with the aforementioned parameters on an NVIDIA H100 PCIe GPU with 80 GB memory. For zero and few-shot demonstration models, we experimented with GPT, specifically *gpt-3.5-turbo-16k and gpt-4o* via OpenAI's Chat Completion API [7] with temperature fixed at 0 for more deterministic behavior. We prompted GPT using the prompts discussed earlier in methodology. For unsupervised methods, we set the summary size (in terms of sentences) for each event to be equal to the number of positive examples in the event from the test set. For LexRank we used cosine similarity threshold 0.1. For BERT-based extractive summarization, we opted for (Miller 2019)'s approach described in the previous section. We used gpt-4o with zero-shots, pairwise ranking strategy and win rate-based aggregation method as an LLM ranker (i.e., extractive summarizer) model. For extractive summarization methods including TextRank, BERT-based summarizer, and LLM ranker, we feed the whole event as context since the existence of the sentences we use for evaluation in the given text is intuitively necessary for these methods in order to prioritize them in the summary.

Results in Figure 3 show variation in performance as the context size increases from 100 to 500 words, with 10-shots GPT-4o as the best performing model. Additionally, few-shot LLM models' performance appears to be more stable with varying context lengths compared to the performance of their zero-shot variants. Moreover, smaller supervised language models show a relatively comparable performance to 10-shot LLMs at context length of 500 words. The majority of these models demonstrated heightened F-1 and accuracy within a context length between 400 and 500 words. This amount corresponds to approximately 12 paragraphs sourced from Reuters.com. This observation indicates that a neutral news report of this specified length pertaining to a given event is sufficient for discerning instances of cherry-picking within a biased narrative. This can be attributed to the common practice in journalism of conveying the important information in the first few paragraphs

[5]https://huggingface.co/google-bert/bert-base-uncased

[6]https://huggingface.co/allenai/longformer-base-4096

[7]https://platform.openai.com/docs/guides/text-generation/chat-completions-api

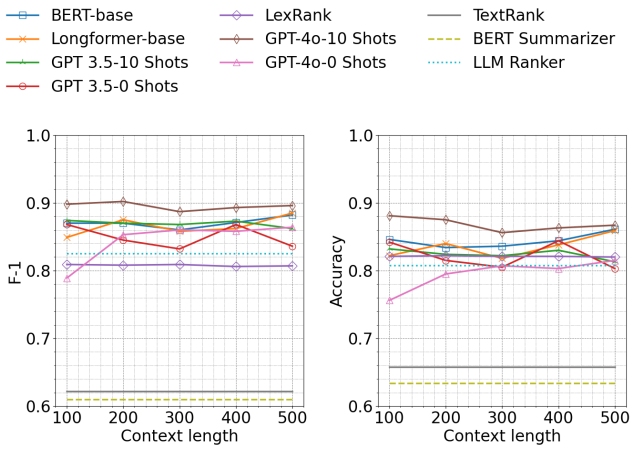

Figure 3: Effect of context size (measured in words) on models' performance. Extractive summarization methods performance is evaluated on full-sized context due to the need for the existence of the sentence in the context in order for it to be extracted in the summary.

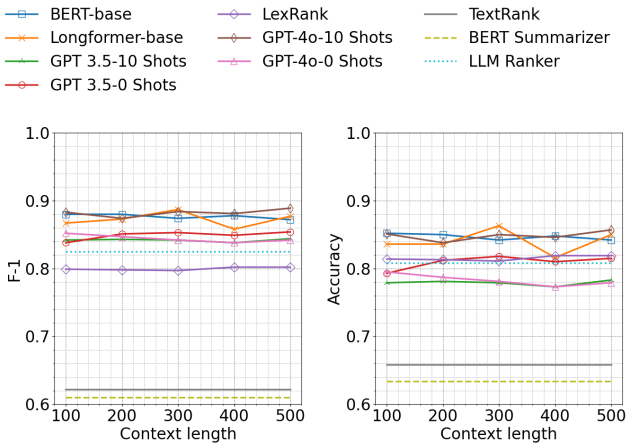

Figure 4: Model performance when using a context collected and summarized in different lengths from biased news sources instead of a neutral source. Extractive summarization methods performance is evaluated on full-sized context due to the need for the existence of the sentence in the context in order for it to be extracted in the summary.

of a news report (Mencher and Shilton 1997). Among summarization methods, LLM ranker and LexRank performed the best. However, summarization methods encounter challenges in capturing important statements when they occur infrequently in the document collection. In the domain of this study, a statement's importance is determined by its critical role as a key piece of evidence, rather than its frequency or similarity to the rest of the text in the news reports.

Regardless of the variation in the performance of the different models, each of the models is still capable of discerning important statements with high accuracy, which provides the flexibility to choose from these models based on the availability of data or computing resources.

(Q.4) Is context derived from sources with counter-biases (i.e., with the absence of a neutral narrative) enough to effectively spot censored missing statements? To answer this question, we experimented with context derived only from biased sources. Particularly, for each event we collected context from a left-biased source and a right-biased source. We then concatenated and fed the two news reports to a generative LLM, specifically GPT 3.5, to summarize them in a single text of varying lengths. The reason for summarizing the two news reports is to make sure their gist fits the upper limits on the input sequence length of the different models. We ran all models against the data sets with summarized contexts (except for extractive summarizers). Additionally, we asked the generative LLM to summarize the context collected from biased sources in 500 words, however, this time we trimmed the 500 words summary at different lengths. We then ran all models against the data sets with summarized-then-trimmed contexts. In these two experiments (i.e., summarized in different lengths vs. summarized at 500 words then trimmed) we used the same settings and hyper-parameters as in the previous experiments. Results in Figure 4 show that our proposed methodology does not rely on the existence of a neutral context to perform well.

Instead, context documents from two sources with different biases can neutralize each other and lead to performance comparable to the least-biased context in the previous experiment. Note how performance becomes more stable when context is summarized then introduced to the models at different lengths as Figure 5 shows. One explanation is that the most important statements are present in the beginning of a news report and thus its summary. This means that once a model sees all the important statements within the context its performance either stops improving or degrades as it gets distracted by additional insignificant text.

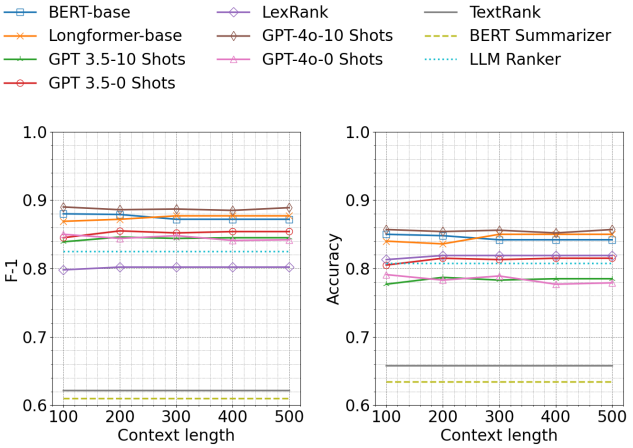

Figure 5: Model performance when using a context collected from biased news sources and summarized in 500 words then trimmed at different lengths. Extractive summarization methods performance is evaluated on full-sized context due to the need for the existence of the sentence in the context in order for it to be extracted in the summary.

(Q.5) Can we mitigate cherry-picking in a biased news story by integrating missing statements from the counter perspective? To answer this question, we used news reports coming from neutral sources as our ground truth d_n from Eq.2. We then randomly selected 100 biased news reports d_i whose semantic similarity with the ground truth news report is less than 0.7, which is the average semantic similarity between all biased reports and their ground truth reports. We then used our best performing supervised model, Longformer-large with 500 words of context and performance of 0.897 and of 0.887 in terms of accuracy and F-1 respectively, to infer the importance scores for all the statements in each event S_e. We filtered out statements (the whole cluster of similar statements) that are not mentioned by the counter perspective sources. This way, we ensured that any statement in S_e is coming only from the counter perspective. Additionally, we ran our statement existence verification LLM skill (refer to Appendix) to filter out statements mentioned by the same source of the biased news report d_i to ensure there is no redundancy in statements in the corrected (i.e., improved) report d_i'. We then identified the missing important statements c_i from each biased document d_i based on Eq.1 and added them to d_i to get d_i'. In total, we collected a sample of 103 triplets (d_i, d_n, d_i'). Finally, we calculated the average cosine similarity between biased reports and their ground truth, and the average cosine similarity between the corrected reports and the ground truth after embedding them using the Sentence Transformer (Reimers and Gurevych 2019) model as in Eq.2. Results for the cosine distances (1-cosine similarity) for each of individual triplet are illustrated in Figure 6. We notice that the majority of corrected news reports are semantically closer to the neutral reports than the biased news reports. The average semantic similarity between corrected reports and neutral reports is 0.695, nearly 12% improvement over the average semantic similarity between the biased (uncorrected) reports and the neutral ones of 0.619. To provide the final draft of the corrected report, missing important statements are integrated within the news report by placing them in their right position within the report. To achieve that, we use an LLM skill that preserves the original text and the smoothness of the flow. For a sample news report corrected by our pipeline, refer to Figure 8 in the Appendix.

We conducted a human evaluation to further assess the readability and bias mitigation of the corrected news articles. Following the Goal Question Metric (GQM) approach (Caldiera and Rombach 1994), we designed evaluation metrics to capture two key aspects: text quality and bias. To construct the text quality evaluation criteria, we referred to the metrics discussed by (Ito, van Deemter, and Suzuki 2025) to create the criteria such as Grammatical Soundness, Narrative Coherence, and Redundancy. For evaluating bias, we adopted questions related to the concept of *card stacking* as described by (Shabo 2008), resulting in criteria such as Balance of Viewpoints, Even-sided Presentation, and Comprehensiveness of Information. To operationalize these criteria, we randomly selected 30 corrected articles and created 30 corresponding surveys for human evaluation. Each criterion was rated using a Likert scale. Notably, the criteria for

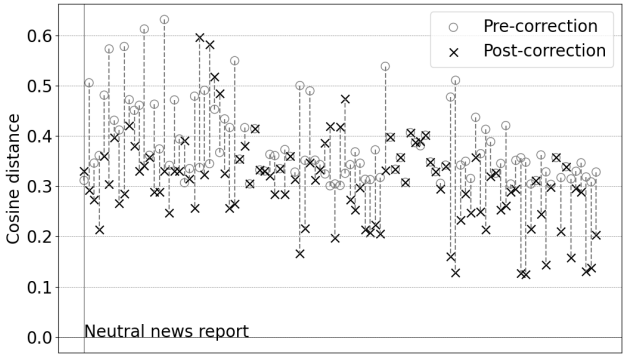

Figure 6: The cosine distances between each triplet of biased, corrected, and ground truth documents.

redundancy and comprehensiveness of information reflect one-directional effects: redundancy ratings are limited to the lower end of the scale (1–3), indicating increasing redundancy, while comprehensiveness is constrained to the upper end (3–5), indicating the degree of information enrichment. Details of the metrics, evaluation criteria, and a sample survey are provided in the Appendix.

A total of nine human evaluators participated in the evaluation. Each evaluator completed 10 surveys, and each survey was assessed by three different evaluators. Specifically, Evaluators 1–3 completed surveys 1–10, Evaluators 4–6 completed surveys 11–20, and Evaluators 7–9 completed surveys 21–30. The average scores for each criterion across survey groups are shown in Table 2. We can see that the newly added statements improved Balance of Viewpoints, Even-Handed Presentation, and Comprehensiveness of Information. They slightly reduced Redundancy, which is a trade-off, and they do not have much impact on Grammatical Soundness and Narrative Coherence. These statistics illustrate the effectiveness of the news articles' correction.

Surveys	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**
0–10	3.1	2.9	2.7	3.3	3.6	4.4
11–20	3.0	2.8	3.0	3.7	3.5	4.0
21–30	2.7	3.2	2.9	4.1	4.1	3.9
All	3.0	3.0	2.9	3.5	3.5	4.1

Table 2: Average scores given by evaluators for different evaluation criteria across survey groups. E1–E6 represent Grammatical Soundness, Narrative Coherence, Redundancy, Balance of Viewpoints, Even-Handed Presentation, Comprehensiveness of Information, respectively.

(Q.6) What levels of a statement's importance granularity can be effectively estimated given the right context? We anticipate that the performance of models trained on varying levels of statement importance granularity will differ. For instance, we expect supervised models to be more precise in drawing a boundary between the labels

Conf. #	Acc.	F-1
1	0.897	0.887
2	0.853	0.828
3	0.875	0.898
4	0.728	0.717

Table 3: Performance of the Longformer model in the four classification configurations.

very important versus everything else, compared to drawing a boundary between more fine-grained importance labels such as *very important* versus *kind of important*. To examine our models' capabilities across different levels of importance granularity, we trained our top-performing model (i.e., Longformer-large) using the four classification configurations outlined in Table 1. The results in Table 3 show that the difficulty of learning a decision boundary decreases when classes are merged and thus label granularity decreases. This trend is particularly notable in the case of *kind of important* and *not very important*. When these two classes are separate (configuration #4 in Table 1), the performance is significantly inferior, compared to other configurations where these two classes are merged. This is attributed to the high fuzziness between the two classes. Therefore, we utilize configuration #1 from Table 1 in all remaining experiments.

(Q.7) Are cherry-picking and general bias correlated in news sources? To answer this question, we examined the relationship between two variables, x: Cherry-picking score calculated on the detected cherry-picking by our automatic pipeline in each news outlet and y: the bias scores of the outlet collected from MBFC and AllSides.

To test the relationship between variables x and y, we ran the full cherry-picking detection pipeline described in Formula 1 employing our top-performing supervised model on 2,453 events comprised of about 97k statements collectively. We ensured that these events did not overlap with the Cherry dataset utilized for training our supervised models. Next, we calculated the average number of cherry-picked statements per outlet across all events, resulting in a final score of x. Finally, since variable y is ordinal, we calculated the Spearman's correlation coefficient r between x and y. Results in Table 4 show a positive moderate correlation approaching statistical significance when the bias scores come from AllSides.com. Additionally, there is a positive weak quasi-significant correlation when the bias scores come from MBFC. [8]

A strong correlation between the two variables is not evident. However, it is important to highlight that the bias scores provided by MBFC and AllSides were calculated by assessing multiple bias forms including those beyond cherry-picking, e.g., lexical cues embedded in text. Furthermore, MBFC and AllSides continually update their bias scores using fresh samples of news stories. On the contrary, the inference of our models was applied on static data from each outlet collected between December 2019 and January 2021.

[8]Further insights into the interpretation of correlation coefficient, including what is considered moderate and what is weak, can be found in (Ratner 2009).

Bias score source	r	P-value
MBFC	0.28	0.10
AllSides	0.32	0.06

Table 4: Correlation between x: cherry-picking detection pipeline scores and y: bias scores from MBFC and AllSides.

Lastly, we hypothesize that the average cherry-picking score x for biased outlets, belonging to a particular bias band, should align with their bias intensity. For instance, we anticipate observing a higher average cherry-picking score among outlets categorized as "left bias" compared to those classified as "left-center bias." To test this, we calculated the mean and standard deviation of the cherry-picking score of all outlets under each bias category as Table 5 shows. Results show an uptrend in cherry-picking scores that aligns with the intensity of bias in the analyzed outlets. Nevertheless, the observed pattern is affected by the limited sample sizes, highlighting the need for future analysis involving larger samples.

Bias category	Mean	STD	Sample size
Left	15.12	4.39	6
Left center	10.32	3.05	15
Right	8.91	4.09	7
Right center	8.33	1.21	5
Center	8.44	0.30	2

Table 5: The mean and standard deviation of cherry-picking scores aggregated by bias category.

Conclusions and Future Work

Manually spotting cherry-picking of statements in news reports is challenging due to the need to examine other narratives to identify missing statements. This study introduces a novel approach to automate the detection and correction of cherry-picking in news reports. Our importance-based approach focuses on comparing multiple news reports that pertain to the same event then identifying and substituting the omission of crucial statements. To facilitate our research, we have constructed the first cherry-picking detection dataset. The results of our models demonstrate promising outcomes. Currently, our work specifically addresses the censorship of statements from a news report. For future work, we plan to study the detection of cherry-picking at larger scales such as selectiveness in events to cover, and subtle scales such as slanting.

Limitations

Our cherry-picking detection approach offers flexibility in selecting a model to assess a statement's importance. However, it also has some limitations, some of which are inherent

to these models, a notable limitation is the use of similarity thresholds. This can be observed in the application of clustering techniques during the creation of events from a set of news reports. If a news source covers an event through multiple reports and one of these reports is not clustered into the correct event, the pipeline may overlook important information from the coverage of that outlet. Additionally, while constructing the data set used in this paper, we clustered statements based on their semantic similarity and a label chosen by the annotator was casted over all statements in the cluster automatically assuming they all convey the same fact. However, in some cases statements that are not completely similar to the rest of the cluster, could be given inaccurate importance label this way. To minimize this, we sacrificed the effort and cost of labeling more examples by setting strict clustering hyper-parameters tuned on a sample of manually labeled 508 statements for this purpose, which led to higher number of clusters of highly similar statements rather than fewer clusters with less similar statements. Moreover, the quality of the evaluation of cherry-picking correction is sensitive to the actual neutrality of the news reports collected from neutral sources since they are used as ground truth in this experiment.

Ethical Statement

While our work focuses on detecting bias by omission, it inevitably incorporates a human element in two stages. First, At the data collection stage, human coders are asked to indicate whether a statement is important to the event. This may result in a dataset that contains the inherent subjectivity and personal perspectives of the human annotators themselves. We are aware that annotators' implicit biases can influence the annotations they provide. Therefore, we ensured that annotators have different political biases in addition to hiding sources names from segments in the annotation tasks, Moreover, we provided annotators with a neutral context to read before they assign labels. Second, bias categorizations we use from MBFC, AllSides, and Ad Fontes Media rely on human annotators as well. Hence, the bias of these sources can be inherently embedded into our dataset. While we hold the belief that the sources we rely on exhibit high quality, it is important to recognize that their reliability and credibility may decline due to various unforeseen factors. At last, we assert that our usage of all artifacts this work including pretrained models (e.g., Longformer and BERT) and derived data (e.g., GDELT events data and labels from MBFC, Ad Fontes Media and AllSides) is purely for research purposes and consistent with their intended use.

References

Asudeh, A.; Jagadish, H. V.; Wu, Y.; and Yu, C. 2020. On detecting cherry-picked trendlines. *Proceedings of the VLDB Endowment*, 13(6): 939–952.

Baly, R.; Da San Martino, G.; Glass, J.; and Nakov, P. 2020. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4982–4991. Online: Association for Computational Linguistics.

Baron, D. P. 2006. Persistent media bias. *Journal of Public Economics*, 90(1-2): 1–36.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Best, J. 2004. *More damned lies and statistics: How numbers confuse public issues*. University of California Press.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Brown, J. A. C. 1963. *Techniques of Persuasion, from propaganda to brainwashing*, volume 604.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Caldiera, V. R. B. G.; and Rombach, H. D. 1994. The goal question metric approach. *Encyclopedia of software engineering*, 528–532.

Chen, W.-F.; Al Khatib, K.; Wachsmuth, H.; and Stein, B. 2020. Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 149–154. Online: Association for Computational Linguistics.

Chiang, C.-F.; and Knight, B. 2011. Media bias and influence: Evidence from newspaper endorsements. *The Review of economic studies*, 78(3): 795–820.

Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C. M.; Brugnoli, E.; Schmidt, A. L.; Zola, P.; Zollo, F.; and Scala, A. 2020. The COVID-19 social media infodemic. *Scientific reports*, 10(1): 1–10.

Dan, V.; Paris, B.; Donovan, J.; Hameleers, M.; Roozenbeek, J.; van der Linden, S.; and von Sikorski, C. 2021. Visual mis-and disinformation, social media, and democracy. *Journalism & Mass Communication Quarterly*, 98(3): 641–664.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Druckman, J. N.; and Parkin, M. 2005. The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, 67(4): 1030–1049.

Erkan, G.; and Radev, D. R. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22: 457–479.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large

spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, volume 96, 226–231.

Fan, L.; White, M.; Sharma, E.; Su, R.; Choubey, P. K.; Huang, R.; and Wang, L. 2019. In Plain Sight: Media Bias Through the Lens of Factual Reporting. 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)*, 6343–6349. Hong Kong, China: Association for Computational Linguistics.

Fleming, C. A. 1995. Understanding propaganda from a general semantics perspective. *ETC.: A Review of General Semantics*, 52(1): 3–13.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gläßel, C.; and Paula, K. 2020. Sometimes less is more: Censorship, news falsification, and disapproval in 1989 East Germany. *American Journal of Political Science*, 64(3): 682–698.

Hamborg, F. 2020. Media Bias, the Social Sciences, and NLP: Automating Frame Analyses to Identify Bias by Word Choice and Labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 79–87. Online: Association for Computational Linguistics.

Hameleers, M.; Brosius, A.; and de Vreese, C. H. 2022. Whom to trust? Media exposure patterns of citizens with perceptions of misinformation and disinformation related to the news media. *European Journal of Communication*, 37(3): 237–268.

Horych, T.; Mandl, C.; Ruas, T.; Greiner-Petter, A.; Gipp, B.; Aizawa, A.; and Spinde, T. 2024. The Promises and Pitfalls of LLM Annotations in Dataset Labeling: a Case Study on Media Bias Detection. *arXiv preprint arXiv:2411.11081*.

Hube, C.; and Fetahu, B. 2018. Detecting biased statements in wikipedia. In *Companion proceedings of the the web conference 2018*, 1779–1786.

Ito, T.; van Deemter, K.; and Suzuki, J. 2025. Reference-free Evaluation Metrics for Text Generation: A Survey. *arXiv preprint arXiv:2501.12011*.

Lee, A.; and Lee, E. B. 1939. *The Fine Art of Propaganda; A Study of Father Coughlin's Speeches, Edited by Alfred Mcclung Lee & Elizabeth Brian Lee*.

Lee, A. M. 1945. The analysis of propaganda: a clinical summary. *American Journal of Sociology*, 51(2): 126–135.

Lee, S.; Gil de Zúñiga, H.; and Munger, K. 2023. Antecedents and consequences of fake news exposure: a two-panel study on how news use and different indicators of fake news exposure affect media trust. *Human Communication Research*.

Lee, T.-T. 2010. Why they don't trust the media: An examination of factors predicting trust. *American behavioral scientist*, 54(1): 8–21.

Leetaru, K.; and Schrodt, P. A. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, 1–49. Citeseer.

Lei, Y.; Huang, R.; Wang, L.; and Beauchamp, N. 2022. Sentence-level Media Bias Analysis Informed by Discourse Structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10040–10050. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Lin, L.; Wang, L.; Zhao, X.; Li, J.; and Wong, K.-F. 2024. IndiVec: An Exploration of Leveraging Large Language Models for Media Bias Detection with Fine-Grained Bias Indicators. *ArXiv*, abs/2402.00345.

Lin, Y.; Youngmann, B.; Moskovitch, Y.; Jagadish, H.; and Milo, T. 2021. On detecting cherry-picked generalizations. *Proceedings of the VLDB Endowment*, 15(1): 59–71.

Mencher, M.; and Shilton, W. P. 1997. *News reporting and writing*. Brown & Benchmark Publishers Madison, WI.

Mihalcea, R.; and Tarau, P. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 404–411.

Miller, D. 2019. Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Paul, R.; and Elder, L. 2008a. *How to Detect Media Bias & Propaganda: In the National and World News: Based on the Critical Conepts & Tools*.

Paul, R.; and Elder, L. 2008b. *How to Detect Media Bias & Propaganda: In the National and World News: Based on the Critical Conepts & Tools*. Foundation for Critical Thinking.

Paul, R.; and Elder, L. 2019. *The Thinker's Guide for Conscientious Citizens on How to Detect Media Bias and Propaganda in National and World News: Based on Critical Thinking Concepts and Tools*.

Podesta, D. 2009. Soft Censorship: How Governments Around the Globe. *A Report to the Center for International Media Assistance*.

Ratner, B. 2009. The correlation coefficient: Its values range between+ 1/- 1, or do they? *Journal of targeting, measurement and analysis for marketing*, 17(2): 139–142.

Recasens, M.; Danescu-Niculescu-Mizil, C.; and Jurafsky, D. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1650–1659.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ribeiro, F.; Henrique, L.; Benevenuto, F.; Chakraborty, A.; Kulshrestha, J.; Babaei, M.; and Gummadi, K. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Rodrigo-Ginés, F.-J.; Carrillo-de Albornoz, J.; and Plaza, L. 2024. A systematic review on media bias detection: What is

media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237: 121641.

Shabo, M. 2008. *Techniques of propaganda and persuasion*. Prestwick House Inc.

Shah, B. S.; Shah, D. S.; and Attar, V. 2024. Decoding News Bias: Multi Bias Detection in News Articles. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, 97–104.

Spinde, T.; Plank, M.; Krieger, J.-D.; Ruas, T.; Gipp, B.; and Aizawa, A. 2021a. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1166–1177. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Spinde, T.; Rudnitckaia, L.; Mitrović, J.; Hamborg, F.; Granitzer, M.; Gipp, B.; and Donnay, K. 2021b. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3): 102505.

Sproule, J. M. 2001. Authorship and origins of the seven propaganda devices: A research note. *Rhetoric & Public Affairs*, 4(1): 135–143.

Strömbäck, J.; Tsfati, Y.; Boomgaarden, H.; Damstra, A.; Lindgren, E.; Vliegenthart, R.; and Lindholm, T. 2020. News media trust and its impact on media use: Toward a framework for future research. *Annals of the International Communication Association*, 44(2): 139–156.

Vincent, R. C. 2019. *Global communication and propaganda*. Rowman & Littlefield Lanham, MD.

Wang, J. S.; Haider, S.; Tohidi, A.; Gupta, A.; Zhang, Y.; Callison-Burch, C.; Rothschild, D.; and Watts, D. J. 2025. Media Bias Detector: Designing and Implementing a Tool for Real-Time Selection and Framing Bias Analysis in News Coverage. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Weld, G.; Glenski, M.; and Althoff, T. 2021. Political bias and factualness in news sharing across more than 100,000 online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 796–807.

Wu, Y.; Agarwal, P. K.; Li, C.; Yang, J.; and Yu, C. 2017. Computational fact checking through query perturbations. *ACM Transactions on Database Systems (TODS)*, 42(1): 1–41.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see the Dataset section.
   (e) Did you describe the limitations of your work? Yes, see the Limitations section.
   (f) Did you discuss any potential negative societal impacts of your work? NA
   (g) Did you discuss any potential misuse of your work? NA
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Introduction and Experimentation and Results sections.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes, see the Experimentation and Results.
   (b) Have you provided justifications for all theoretical results? Yes, see the Experimentation and Results.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, see the Related Work.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, see the Experimentation and Results.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes, see the Limitations.
   (f) Have you related your theoretical results to the existing literature in social science? Yes, see the Experimentation and Results.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, link to our anonymous repository is provided throught the paper for data, code, and experiments.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, See Experimentation and Results.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, because variations are not significant to the task the paper addressing.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see the Experimentation and Results.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see the Experimentation and Results.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes. See Limitations.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes.

(b) Did you mention the license of the assets? Yes. See section Introduction

(c) Did you include any new assets in the supplemental material or as a URL? Yes.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA. Our data is published news reports.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA. The data consists of published (publicly available) news articles and does not contain any sensitive information types or offensive content.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA.

(d) Did you discuss how data is stored, shared, and deidentified? NA.

Appendix

A. Cherry-picking of Facts in News Reporting - Example

Table 6 illustrates cherry-picking in journalism by showcasing statements from authentic news reports by distinctively biased news outlets, all covering the same event: U.S. President Donald Trump's attendance at the annual Davos economic forum. The three news outlets included in Table 6 are Reuters.com (center), NewsMax.com (right-bias), and CNBC.com (left-bias). The bias categorization of these media outlets is determined by assessments from Media Bias Fact Check (MBFC). The table organizes corresponding statements from the three news reports in rows based on their semantic similarity. If a statement is not mentioned by a particular news source, the corresponding cell in the table is left empty. Note that Reuters.com and NewsMax.com published identical reports for this event. By examining Table 6, we can observe instances of cherry-picking. For example, in the third row, two of the news outlets mentioned that President Trump attended the Davos forum last year (i.e., 2019). However, CNBC.com (left-bias) failed to mention it.

B. Media Bias Rating Sources

MBFC, AllSides, and Ad Fontes Media employ diverse methods to evaluate bias in news outlets. Literature addressing bias and misinformation detection heavily relies on these three sources for bias categorization in tasks such as social media source annotation (Weld, Glenski, and Althoff 2021), comparing the diffusion of news from reliable and non-reliable sources (Cinelli et al. 2020), and benchmarking automatic media bias monitors (Ribeiro et al. 2018).

Media Bias Fact Check (MBFC) MBFC utilizes a comprehensive methodology to assess the ideological leanings and factual accuracy of media outlets.[9] The bias assessment includes categories such as Biased Wording/Headlines, Factual/Sourcing, Story Choices, and Political Affiliation. The scoring mechanism categorizes sources into Least Biased, Left/Right Center Bias, Left/Right Bias, or Extreme Bias. The methodology also assesses factual reporting with ratings ranging from Very High to Very Low based on the reliability and commitment to factual accuracy.

AllSides AllSides employs a comprehensive methodology to rate media bias, including various factors and perspectives.[10] AllSides Editorial Reviews involve a multipartisan panel assessing news reports and bias indicators. Blind Bias Surveys gather opinions from average Americans across political spectrums. AllSides reviewers assess content independently, considering common bias indicators and transparency about political leaning. AllSides also incorporates third-party data and allows Community Feedback for additional perspectives.

Ad Fontes Media Ad Fontes Media generates scores for news sources based on the ratings of individual articles or episodes, using over 60 trained analysts from diverse backgrounds who undergo initial and ongoing training. [11] Each source is rated by at least three analysts with different political leanings. The analysts independently rate the content,

[9] https://mediabiasfactcheck.com/methodology/

[10] https://www.allsides.com/media-bias/media-bias-rating-methods

[11] https://adfontesmedia.com/how-ad-fontes-ranks-news-sources/

Reuters.com (Center) / NewsMax.com (Right-bias)	CNBC.com (Left-bias)
U.S. President Donald Trump plans to attend the annual Davos economic forum in January.	After skipping a gathering of the world's most elite in Davos, Switzerland last January, President Donald Trump will attend the World Economic Forum in 2020.
Trump had to cancel his plan to attend the annual gathering of global economic leaders early this year due to a government shutdown.	Trump blamed his no-show last time on the partial government shutdown that was triggered by a funding dispute over a proposed wall along the United States' southern border.
He attended the Davos forum last year.	-
The exact date of when Trump would participate was unclear.	A spokesperson for the White House did not immediately respond to a request for comment about the president's plans to attend next year.
Davos may be one of the few foreign trips that Trump takes in 2020.	-
The World Economic Forum in the Swiss ski resort town of Davos is scheduled to run January 21-24.	The 50th annual forum, held Jan. 21 - Jan. 24 will focus on "Stakeholders for a Cohesive and Sustainable World."
-	Meantime, while the Davos agenda will look to "create bridges to resolve conflicts in global hotspots," Trump has made jabs at a number of European allies.
-	While he has said he was just joking about placing tariffs on cars imported from the European Union, he is still threatening to place tariffs on French specialties like champagne and cheese.

Table 6: A juxtaposition of three news stories covering the same event, published by three sources with different political biases. Bias categorization of these media outlets is based on MediaBiasFactCheck.com (MBFC).

compare scores, and discuss any discrepancies. The overall rating is the average of the analysts' ratings. In some cases, more analyses may be used for rating articles with outlier scores.

C. Data Annotation Interface

To annotate the dataset for the purpose of this study, we developed a tool that meets our needs. The tool shows the annotators a context, followed by statements clustered based on similarity. Annotators are required to indicate the importance of the statements shown after reading the context as shown in a screenshot from the tool in Figure 7.

D. Clustering of S_e

Table 7 shows two statement clusters from two different events.

E. Global attention in Longformer

To evaluate the influence of the global attention mechanism on predicting a statement's importance, we tested our best model (Longformer-large) with all possible combinations of global attention locations within the input sequence, i.e., the classification token, the statement tokens, and the context tokens. We maintained the sequence length of 512 and set the batch size to 4 in this experiment to ensure memory can handle applying global attention on all the sequence tokens. Results in Table 8 show that we can still maintain the same performance but optimize memory and inference time (Beltagy, Peters, and Cohan 2020) by applying global attention on the statement tokens and classification tokens only. For this reason, we fixed global attention to these two locations in all of our experiments. Moreover, we can optimize further and trade off more efficiency at the expense of a slight decrease in performance by applying global attention on the classification token only.

F. Evaluation Metrics of News Articles

To evaluate the effectiveness and quality of neutralized news reports produced by our cherry-picking correction pipeline, we utilize the following metrics in the user study we conducted for this purpose:

Fluency. Assess whether the news article maintains grammatical fluency, logical organization, and an easy-to-follow narrative.

Purpose: Ensure that the news article maintains a fluent, logically organized, and easy-to-read narrative. The article should be grammatically sound, flow smoothly, and avoid unnecessary repetition that could disrupt comprehension.

Evaluation Criteria:

- Grammatical Soundness: Sentences are grammatically correct and sound natural.
- Narrative Coherence: Ideas are logically connected and the article maintains a consistent narrative flow.
- Redundancy: No unnecessary repetition of ideas or statements.

Bias and Framing. Assess whether the news article minimizes any bias in its framing and presentation of information and more comprehensive and neutral in terms of presenting the different perspectives.

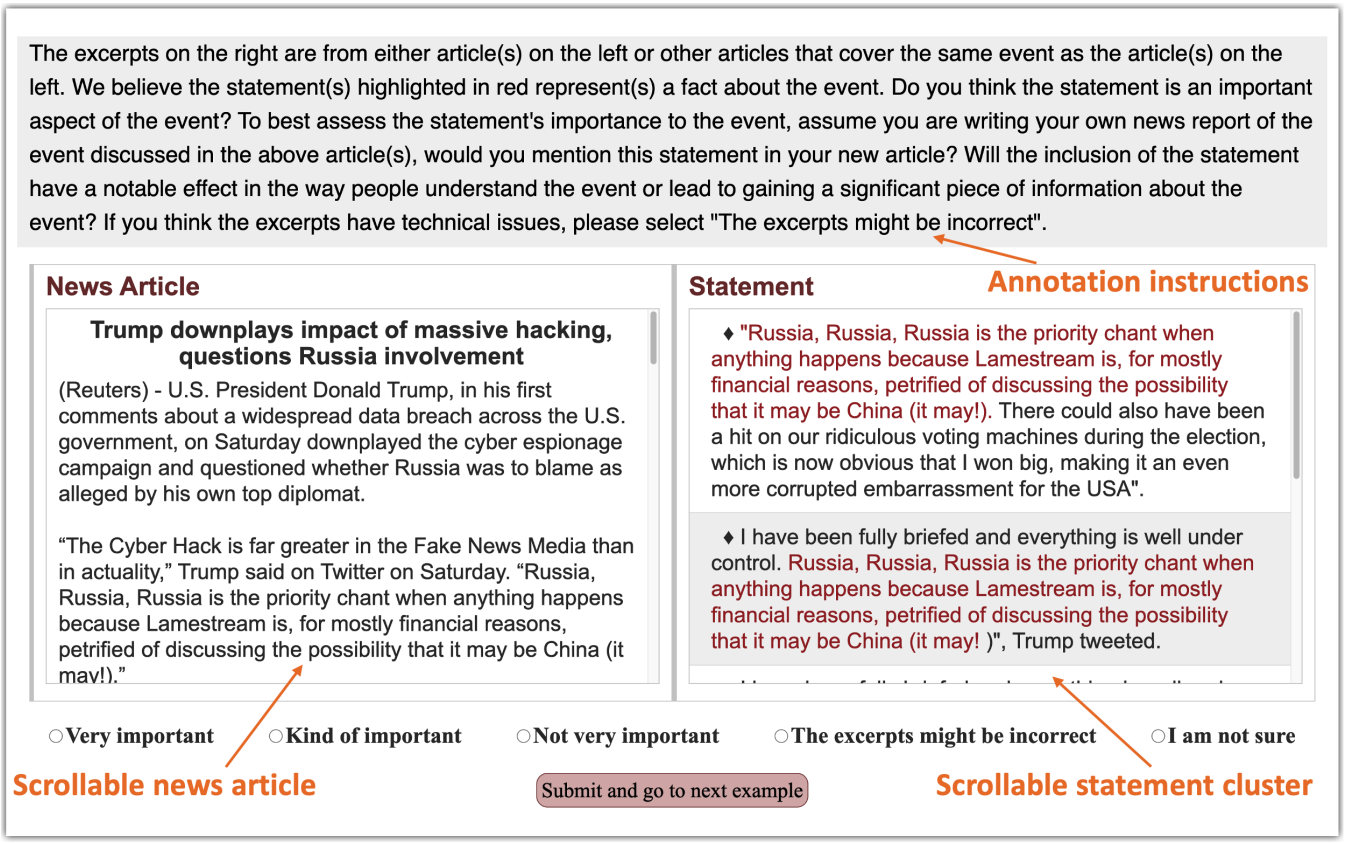

Figure 7: Cherry-picking data annotation interface.

Evaluation Criteria:

- Balance of Viewpoints: The article acknowledges multiple sides fairly including the opposing perspectives, rather than promoting one viewpoint disproportionately
- Even-Handed Presentation: No side is presented with significantly more evidence, explanation, or emphasis than others.
- Comprehensiveness of Information: The article is thorough and comprehensive in covering key facts, arguments, and contextual information related to the topic.

G. User Study Survey Example

Instruction:

Please read the following news reports and answer the questions about how adding the highlighted sentences affects the original report (black text).

Article:

BAGHDAD — President Donald Trump is blaming Iran for a breach of the U.S. Embassy compound in Baghdad and is calling on Iraq to protect the embassy. Trump tweeted Tuesday that "Iran killed an American contractor, wounding many." **The US carried out five airstrikes in Iraq and Syria on Sunday on facilities controlled by Kataib Hezbollah, killing at least 25 people and wounding 51, in the first significant US military response to weeks of deadly rocket attacks by the Iran-backed group on US-Iraqi targets. The American airstrikes killed 24 members of an Iranian-backed militia, the Kataeb Hezbollah, in retaliation for last week's killing of an American contractor in a rocket attack on an Iraqi military base.** Members of the Hashed al-Shaabi, a mostly Shiite network of local armed groups trained and armed by powerful neighbor Iran, smashed the bullet-proof glass of the US embassy's windows in Baghdad with blocks of cement after breaching the outer wall of the diplomatic mission on December 31, 2019, to vent their anger over the weekend airstrikes that killed pro-Iran fighters in western Iraq. Trump says, "We strongly responded, and always will. Now Iran is orchestrating an attack on the U.S. Embassy in Iraq. They will be held fully responsible. In addition, we expect Iraq to use its forces to protect the Embassy, and so notified! Trump tweeted from his estate in Palm Beach, Florida, where he is in the midst of two-week plus vacation. He's been largely out of sight and the tweet marked his first comment on the weekend U.S. airstrikes in Iraq and Syria.

Questions:

1. **Grammatical Soundness**: Do the highlighted sentences improve or reduce the grammatical correctness and naturalness of the article?

- Greatly improve
- Somewhat improve

Cluster #1
· President-elect Joe Biden plans to release nearly all available doses of the COVID-19 vaccines after he takes office. · President-elect Joe Biden plans to release almost all vaccine doses immediately. · President-elect Joe Biden will aim to release every available dose of the coronavirus vaccine when he takes office. · Joe Biden will release most available Covid-19 vaccine doses to speed delivery to more people when he takes office.
Cluster #2
· The House voted to override President Trump's veto of a $740 billion defense spending and policy bill. · The House of Representatives on Monday voted to override President Trump's veto of the National Defense Authorization Act for Fiscal Year 2021. · House Votes to Override Trump's Veto of 2021 Defense Policy Bill. · The House voted late Monday to override President Donald Trump's veto of a defense spending bill for 2021.

Table 7: Two examples of statement clusters.

Global attention locations	**Acc.**	**F-1**
[CLS]	0.819	0.820
[CLS] + statement	0.831	0.837
[CLS] + context	0.683	0.680
[CLS] + statement + context	0.837	0.838

Table 8: Performance of the Longformer-large model using different global attention locations.

New York City is taking extra measures to protect major areas from any potential revenge attacks after a US airstrike killed a top Iranian general.
Mayor Bill de Blasio said he's spoken with top city officials about immediate steps the New York Police Department will take to protect key locations "from any attempt by Iran or its terrorist allies to retaliate against America." The city will be "vigilant against this threat for a long time to come," de Blasio tweeted.
The Los Angeles Police Department said there are no credible threats to the city either but it's monitoring the developments in Iran.
"We will continue to communicate with state, local, federal and international law enforcement partners regarding any significant intel that may develop," it tweeted.
The US airstrike that killed Iran Quds Force commander Qasem Soleimani generated starkly different reactions along party lines Thursday night. Soleimani was the leader of the Quds force, a division of the IRGC trained in unconventional warfare beyond Iran's borders, including Syria and Iraq. Defense officials said Soleimani was planning to attack U.S. diplomats and service members throughout the region. Republicans heaped praise on President Donald Trump while Democrats expressed concerns about the legality and consequences of the attack.

Figure 8: A news report corrected by integrating the two underlined missing important statements in their right position within the report.

- No impact
- Somewhat reduce
- Greatly reduce

2. **Narrative Coherence**: Do the highlighted sentences enhance or disrupt the logical flow and consistency of the article's narrative?

- Greatly enhance
- Somewhat enhance
- No impact
- Somewhat disrupt
- Greatly disrupt

3. **Redundancy**: Do the highlighted sentences introduce redundant content?

- No impact
- Somewhat increase redundancy
- Greatly increase redundancy

4. **Balance of Viewpoints**: Do the highlighted sentences contribute to a more balanced representation of opposing viewpoints, or do they skew the coverage?

- Greatly improve balance
- Somewhat improve balance
- No impact
- Somewhat reduce balance
- Greatly reduce balance

5. **Even-Handed Presentation**: Do the highlighted sentences help ensure both sides of the issue are covered with similar depth and attention?

- Greatly improve even-handedness
- Somewhat improve even-handedness
- No impact
- Somewhat reduce even-handedness
- Greatly reduce even-handedness

6. **Comprehensiveness of Information**: Do the highlighted sentences add important facts or perspectives that improve the overall coverage of the topic?

- Greatly improve comprehensiveness
- Somewhat improve comprehensiveness
- No impact

G. LLM Templates

Importance classification We experimented with various prompts for importance classification task, including simple ones such as the question *"Is the above sentence important to mention in a news article that covers the story mentioned in the above news article? Answer with 'yes' or 'no' only."* Using the prompt as shown in Figure 9, which is crafted based on the annotation interface, yielded the best GPT-based performance on the test dataset.

```
##You are a journalist. In this exercise:
You will first read the following news
article about an event. This is a real
news article. It is from coverage of
events during the period of 12/15/2019 to
01/08/2021. You will be given a statement
about the event that may be discussed in
the article you read or other related
articles which you didn't read. You
will be asked to assess the statement's
importance with regard to the event.
...
10-shots ...
...
Task #n:
ARTICLE:
{context}
QUESTION: The following statement is
from the above article or other articles
that cover the same event as the above
article. Do you think the statement is
an important aspect of the event? To best
assess the statement's importance to the
event, assume you are writing your own
news report of the event discussed in
the above article, would you mention this
statement in your new article? Will the
inclusion of the statement have a notable
effect in the way people understand the
event or lead to gaining a significant
piece of information about the event?
Answer with "yes" or "no" only.
STATEMENT: {sentence}
```

Figure 9: LLM prompt template for the importance classification task.

Statement existence verification We use the prompt in 10 to verify if a given statement exists in a given news report.

```
## You are a helpful journalist.
## You are given:
1. A statement that contains an
information or a fact.
2. A news report
## Your task: Read the news report,
and verify if the information in the
statement exists in the news report in
a way or another.
## Your output: Answer **only** with "1"
if the information in the statement
exists in the news report, and "0"
otherwise.
STATEMENT: {statement}
NEWS REPORT: {news report}
```

Figure 10: LLM skill for verifying the existence of a statement in a news report.

Missing statement integration We use the prompt in 11 to integrate a given missing statement within a given news report while preserving flow and style of the original text. Figure 8 shows a sample news report after correcting it by integrating missing important statement using the integration skill.

```
## You are a helpful journalist.
## You are given:
1. A statement that contains a fact or
information about an event.
2. A news report that covered the same
event.
## Your task: Read the news report, and
integrate the statement within the news
report. Place the statement in its most
suitable place within the news report to
guarantee a smooth flow of information.
Try to not change the style of writing or
the words and terms as much as possible,
UNLESS NEEDED.
## Your output: Return the full rewritten
news report with the statement integrated
in it. DO NOT return or answer with
anything else!!!
STATEMENT: {statement}
NEWS REPORT: {news report}
REWRITTEN NEWS REPORT:
```

Figure 11: LLM skill for integrating a missing statement within a news report.