Multi-document Summarization through Multi-document Event Relation Graph Reasoning in LLMs: a case study in Framing Bias Mitigation

Yuanyuan Lei and Ruihong Huang
Department of Computer Science and Engineering
Texas A&M University, College Station, TX
{yuanyuan, huangrh}@tamu.edu

Abstract

Media outlets are becoming more partisan and polarized nowadays. Most previous work focused on detecting media bias. In this paper, we aim to mitigate media bias by generating a neutralized summary given multiple articles presenting different ideological views. Motivated by the critical role of events and event relations in media bias detection, we propose to increase awareness of bias in LLMs via multi-document events reasoning and use a multi-document event relation graph to guide the summarization process. This graph contains rich event information useful to reveal bias: four common types of in-doc event relations to reflect content framing bias, cross-doc event coreference relation to reveal content selection bias, and event-level moral opinions to highlight opinionated framing bias. We further develop two strategies to incorporate the multi-document event relation graph for neutralized summarization. Firstly, we convert a graph into natural language descriptions and feed the textualized graph into LLMs as a part of a hard text prompt. Secondly, we encode the graph with graph attention network and insert the graph embedding into LLMs as a soft prompt. Both automatic evaluation and human evaluation confirm that our approach effectively mitigates both lexical and informational media bias, and meanwhile improves content preservation[1].

1 Introduction

Media bias refers to the practice of presenting biased or partial information in news articles to promote an ideological leaning and sway readers' political opinions (Gentzkow and Shapiro, 2006; Groeling, 2013; Morstatter et al., 2018). News media plays a crucial role not only in supplying information, but also in selecting and organizing information to shape public opinions (Baron, 2005; Asp, 2007; Hillebrand, 2019; Lei and Huang, 2022; Lei et al., 2024b). With the media outlets become more partisan and polarized, the journalists usually embed their ideological bias into news articles through content framing (Tankard Jr, 2001; Prior, 2013; D'Angelo, 2017). The prevalence of media bias has harmful effects on both individuals and society, such as misleading audiences, intensifying societal polarization, and undermining democratic values (Kuypers, 2002; Druckman and Parkin, 2005; Lei and Cao, 2023; Entman, 2007; Emami et al., 2020).

While media bias presents a significant issue, most previous research focused on detecting media bias and few efforts were made to mitigate media bias (Fan et al., 2019; Baly et al., 2020; Naredla and Adedoyin, 2022; Liu et al., 2022a). Recently, neutralized summarization (Lee et al., 2022) was proposed to mitigate media bias by generating a neutralized summary given multiple articles that frame the same story from liberal or conservative viewpoints. This task holds great promise in offering a comprehensive view of news reporting and enabling unbiased information access. However, the approaches for neutralized summarization remain rudimentary and mainly rely on basic text-to-text generation that may suffer from limited content analysis or lack of awareness of bias (Lee et al., 2022; Bang et al., 2023).

To mitigate media bias, we argue that it is necessary to incorporate bias indicators to inform the neutralized summarization process. Motivated by the critical roles of events and event relations in detecting media bias (Zhang et al., 2021; Liu et al., 2023; Zhang et al., 2024; Lei and Huang, 2024b), we propose to inform LLMs of bias distribution through multi-document events reasoning.

In particular, we propose to build a multi-document event relation graph that takes events as nodes and captures in-doc and cross-doc event relations. This graph contains various bias relevant information: (1) four types of in-doc event

[1]The code and data link is: https://github.com/yuanyuanlei-nlp/multi_doc_summarization_acl_2025

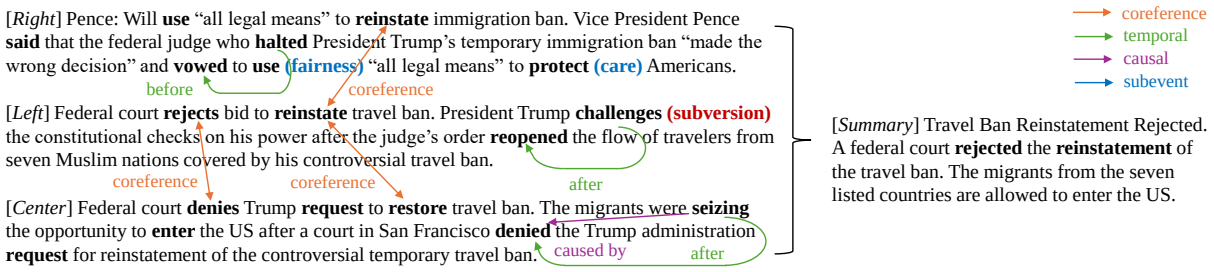


Figure 1: An example of multi-document event relation graph based on a triplet of articles. The multi-document event relation graph consists of events as nodes (bold words), moral opinions as event node attribute (colored words in parentheses), and within-doc and cross-doc event relations (colored edges between events).

relations (temporal, causal, subevent, and coreference) illustrate the diverse narrative logic connecting events within an article, thereby reflecting content framing bias (2) cross-doc event coreference relations distinguish between events commonly reported across multiple articles and events selectively reported by a particular article, thus revealing content selection bias (3) we also add event-level moral opinions as a feature for event nodes to highlight morally opinionated events and indicate moral framing differences across articles. The designed multi-document event relation graph is expected to increase awareness of bias in LLMs and serve as an useful guidance for neutralized summarization.

Take the example in Figure 1 as an illustration, where the three articles report essentially the same news of *Travel Ban Reinstatement*, but from perspectives of different ideological leanings. The bold words are event words and the edges are colored to represent different event relation types. We can see that in this multi-document event relation graph, the cross-doc event coreference relations highlight common events shared across the articles, and the main *Reinstatement* event appeared in all the articles. Then, among the remaining events unique to each article, the right leaning article chooses to frame the main event as *use "all legal means"* and describe the purpose as *protect Americans*, while the left leaning article frames the main event as *challenges the constitutional checks*. Further, moral sentiment analysis shows that *use "all legal means"* and *protect Americans* have positive moral sentiments of fairness and care respectively, while *challenges the constitutional checks* has a negative moral sentiment of subversion.

We further propose a framework to integrate the multi-document event relation graph into LLMs for neutralized summarization, which consists of two key components. The first is graph textualization, where we convert the multi-doc event relation graph into natural language descriptions, and feed the textualized graph as a hard prompt into LLMs. The second is graph prompt tuning, where we encode the multi-doc event relation graph with graph attention network, and insert the graph embedding as a soft prompt into LLMs for tuning. The incorporation of hard and soft prompt are complementary: the hard prompt informs the model of graph structure by augmenting the instruction, while the soft prompt enables direct tuning on graph embedding. Both automatic evaluation and human evaluation confirm the effectiveness of our approach based on multi-doc event relation graph, which notably mitigates both lexical and informational media bias in summaries and meanwhile improves the preservation of content semantics. Our main contributions are summarized as follows:

- We propose to incorporate bias relevant information for media bias mitigation through multi-document events reasoning.
- We introduce a multi-document event relation graph to guide neutralized summarization.
- We design a new framework to integrate the graph into LLMs, reducing media bias while also improving content preservation.

2 Related Work

Multi-document Summarization aims to generate a concise and informative summary from a collection of documents (Lebanoff et al., 2018). In recent years, researchers applied deep neural networks and large pre-trained language models for multi-document summarization (Mao et al., 2020; Pasunuru et al., 2021; Shen et al., 2023). Additionally, researchers explored several subtopics in this field, such as topic-guided, agreement-oriented,

or entity-aware summarization (Cui and Hu, 2021; Pang et al., 2021; Zhou et al., 2021). Differently, our goal is to generate a neutralized and unbiased summary from multiple articles with varying ideology, thereby mitigating framing bias.

Event Graph was introduced by Li et al. (2020); Jin et al. (2022), which includes entity-entity links and event-entity links via event argument roles, yet lacks event-event relations. This graph is employed in several down-stream tasks, including story generation (Chen et al., 2021), misinformation detection (Wu et al., 2022), and sentence fusion (Yuan et al., 2021). In contrast, our graph focuses on events and establishes interrelations between events through four types of event-event relations.

Event Relations were studied for decades. There are four common relations between events: coreference, temporal, causal, and subevent relations (Caselli and Vossen, 2017; Zeng et al., 2020; Tan et al., 2021; Man et al., 2022; Lai et al., 2022; Wang et al., 2022; Lei and Huang, 2023b). While each type of relations was previously studied in isolation, we aim to develop a model that unifies all four relations for comprehensive content analysis. Instead of analyzing event relations in single article, we propose to construct a multi-document event relation graph to capture narrative structures across various articles.

Media Bias Detection attracted research interests for years (Lichter, 2017). Early work detect media bias at source level, by assuming all the articles within one media source share the same ideology (Budak et al., 2016; Baly et al., 2018). Subsequent research shifted towards detecting media bias at article level, by classifying the ideology leaning of each article (Sapiro-Gheiler, 2019; Baly et al., 2020; Chen et al., 2020; Liu et al., 2022b). More recently, there has been an interest in detecting media bias at more granular levels, such as sentence level or token level (Da San Martino et al., 2019; van den Berg and Markert, 2020; Spinde et al., 2021; Vargas et al., 2023; Lei et al., 2022; Lei and Huang, 2023a, 2024a). Different from most previous work that develop approaches for media bias detection, this paper aims for media bias mitigation.

Media Bias Mitigation has a relatively short research history. The first work for mitigating media framing bias was introduced by Lee et al. (2022), where they introduced the NeuS dataset for neutralized summarization. They also designed a method to generate summary in a hierarchical order from title to article (Lee et al., 2022). Different from previous work based on text-to-text model, our work firstly incorporates bias information into this task. We propose a multi-document event relation graph approach to inform LLMs of bias distribution and guide LLMs in mitigating framing bias.

3 Multi-document Event Relation Graph

Given a cluster of news articles, we propose to create a multi-document event relation graph for content analysis. Overall, the graph comprises events as nodes, moral opinions as event node attributes, four common types of single-doc event relations, as well as cross-doc event coreference relation to connect articles together.

3.1 Event and Moral Attributes

An *event* refers to an occurrence or action reported in news articles, and is the basic element in story telling (O'Gorman et al., 2016). In news media, the authors often convey their political stance through moral judgment towards events, evaluating whether the events align with social moral rules. The sociologists have developed Moral Foundation Theory to categorize social moral rules into five dimensions, each associated with a positive and negative judgment: *Care / Harm*, *Fairness / Cheating*, *Loyalty / Betrayal*, *Authority / Subversion*, and *Purity / Degradation* (Graham et al., 2009).

The first step of graph construction is extracting events from each article. An event identification model is trained on the MAVEN dataset which annotated event mentions for general-domain documents (Wang et al., 2020) (details in Appendix A). Given a candidate article consisting of N words, we infer the trained event identifier to predict the probability of each word triggering an event:

$$P_i^{event} = (p_i^{event}, p_i^{non-event}) \quad (1)$$

Subsequently, we extract the moral opinion towards each event as node attribute. A moral classifier is trained based on the EMONA dataset which annotated event-level moral opinions in news articles (Lei et al., 2024a). For all the extracted events, we use the moral classifier to predict their moral judgments into ten moral values or non-moral class:

$$P_i^{moral} = (p_i^{care}, p_i^{harm}, ..., p_i^{non-moral}) \quad (2)$$

3.2 Event-Event Relations

There are four common event relations. Coreference relation informs us whether the two events

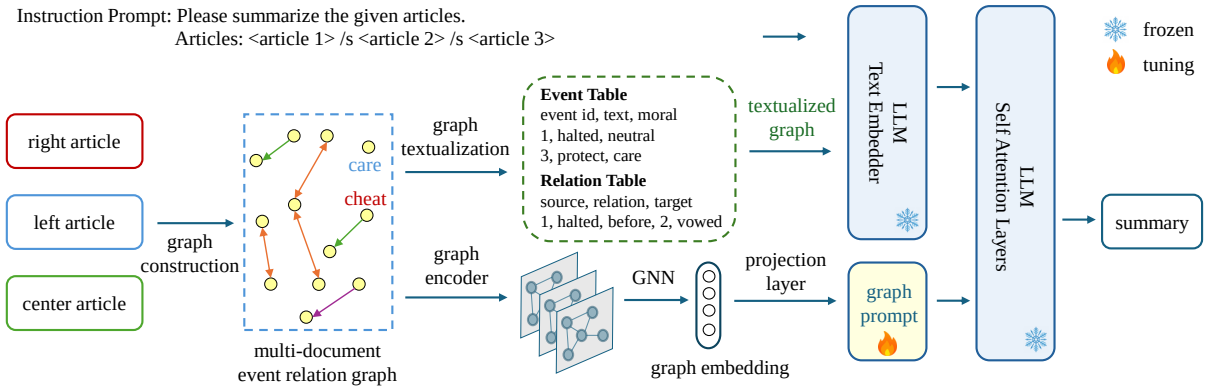


Figure 2: An illustration of neutralized summarization guided by multi-document event relation graph.

designate the same occurrence or not. Temporal relation represents the chronological orders between events, such as *before*, *after*, and *overlap*. Causal relation shows the causality or precondition relation between events, and is categorized into *causes* and *caused by*. Subevent relation recognizes containment or subordination relation between events, including *contains* and *contained by* classes.

The next step is connecting the events with four event relations for each article. The four event relation extractors are trained on the general-domain MAVEN-ERE dataset (Wang et al., 2022). Since the four event relations interact with each other to form a cohesive narrative structure, we adopt the joint learning framework to train these relations collaboratively (Wang et al., 2022). During the inference process, we establish all possible event pairs based on the extracted events. For each event pair $(event_i, event_j)$, we employ the trained relations extractors to predict the probabilities for the four relations. The final label for each relation is derived by applying the $argmax$ function on these predicted probabilities:

$$P_{i,j}^{corefer} = (p_{i,j}^{corefer}, p_{i,j}^{non-corefer}) \quad (3)$$

$$P_{i,j}^{tem} = (p_{i,j}^{before}, p_{i,j}^{after}, p_{i,j}^{overlap}, p_{i,j}^{non-tem}) \quad (4)$$

$$P_{i,j}^{causal} = (p_{i,j}^{causes}, p_{i,j}^{caused-by}, p_{i,j}^{non-causal}) \quad (5)$$

$$P_{i,j}^{sub} = (p_{i,j}^{contains}, p_{i,j}^{contained-by}, p_{i,j}^{non-sub}) \quad (6)$$

3.3 Cross-doc Event Coreference

The final step is connecting the events across different documents through event coreference relation. We employ a cross-document event coreference resolution system (Lai et al., 2021; Wen et al., 2021) to identify clusters of events from multiple documents. The cross-document event coreference relation is used to connect multiple single-document event relation graphs together, facilitating cross-document content analysis and narrative comparison.

4 Neutralized Summarization

The multi-document event relation graph is then incorporated into LLMs for neutralized summarization through two key components (Figure 2). The first is graph textualization, where we convert the graph into natural language descriptions, and feed the textualized graph into LLMs as a hard prompt. The second is graph prompt tuning, where we encode the graph with graph neural network, and insert the graph embedding into LLMs as a soft prompt. The hard and soft prompts complement each other: the hard prompt augments the instruction with graph structure, while the soft prompt enables direct tuning on graph embeddings.

4.1 Graph Textualization

The graph textualization process is designed to transform the graph structure into a natural language format, making it readable by LLMs. This involves creating an event table T_{event} to describe the events information, including event id, event text, and event-level moral judgment. Additionally, a relation table $T_{relation}$ is developed to describe the relations information between events, which includes columns for source event, relation, and target event. The two tables T_{event} and $T_{relation}$ convert the graph structure into textual descriptions, resulting in a textualized graph. This textualized graph is then fed into LLMs as a hard prompt:

$$h_t = TextEmbedder(T_{event}; T_{relation}) \quad (7)$$

where $TextEmbedder$ is the text embedding layer of a pre-trained and frozen LLM.

4.2 Graph Prompt Tuning

The graph prompt tuning process is designed to create a graph embedding and project it as a soft prompt into LLMs for further tuning. This involves a graph propagation process to update events embeddings with their neighbor events embeddings through interconnected relations, and produce a final graph embedding that represents the entire graph. In addition, a projection layer is crafted to transform the graph embedding into the same representation space of LLMs.

During the graph propagation process, we encode the article with Longformer (Beltagy et al., 2020), and use the corresponding word embeddings to initialize event node embeddings e_i. Then, we update event embeddings with their moral values:

$$e_i = W^m(e_i \oplus m_i) + b^m \tag{8}$$

where m_i is the moral label embedding of the event e_i, $\oplus$ denotes feature concatenation, W^m, b^m are trainable parameters of a transformation layer.

Afterwards, we develop a relation-aware graph attention network to update event embeddings with neighbor events embeddings through their linked relations. Given a pair of events (e_i, e_j), their relation r_{ij} is initialized as the embedding of the corresponding relation word. At the l-th layer, the input for i-th event node are output features produced by the previous layer denoted as $e_i^{(l-1)}$. The relation embedding r_{ij} is updated as:

$$r_{ij} = W^r(e_i^{(l-1)} \oplus r_{ij} \oplus e_j^{(l-1)}) \tag{9}$$

where W^r are trainable matrix. Then the attention weights across neighbor events are computed as:

$$\alpha_{ij} = softmax_j\Big((W^Q e_i^{(l-1)})(W^K r_{ij})^T\Big) \tag{10}$$

where W^Q, W^K are trainable parameters. The output feature for e_i regarding the relation r is :

$$e_{i,r}^{(l)} = \sum_{j \in \mathcal{N}_{i,r}} \alpha_{ij} W^V r_{ij} \tag{11}$$

where $\mathcal{N}_{i,r}$ denotes the neighbor event nodes that connect with event e_i via the relation r, and $r \in R = \{$*coreference*, *before*, *after*, *overlap*, *causes*, *caused by*, *contains*, *contained by*$\}$. After collecting $e_{i,r}^{(l)}$ for all relation types R, the final output feature for event e_i at l-th layer is aggregated as:

$$e_i^{(l)} = \sum_{r \in R} e_{i,r}^{(l)} / |R| \tag{12}$$

Subsequently, we derive the graph embedding by introducing an additional graph node and linking it to the event nodes. We employ the standard graph attention network to aggregate event embeddings into the graph embedding h_g^l at l-th layer:

$$\alpha_i = softmax_i\Big(W h_g^{(l-1)} \oplus W e_i^{(l-1)}\Big) \tag{13}$$

$$h_g^{(l)} = \sum_i \alpha_i W e_i^{(l-1)} \tag{14}$$

The graph embedding from the last layer is designated as the final graph embedding h_g.

Furthermore, a projection layer is designed to transform the graph embedding into the same representation space of the LLM:

$$\hat{h}_g = W_2\Big(W_1 h_g + b_1\Big) + b_2 \tag{15}$$

where W_1, W_2, b_1, b_2 are the parameters of the projection layer, and $\hat{h}_g$ is the resulting graph prompt.

During the summarization procedure, both the graph prompt $\hat{h}_g$ and textualized graph h_t are fed into the self attention layers of a pre-trained and frozen LLM. The graph prompt $\hat{h}_g$ receives gradients and enables back propagation.

5 Experiments

5.1 Datasets

The task of neutralized summarization has a relatively short research history, and NeuS (Lee et al., 2022) is the only available dataset up till now.

NeuS (Lee et al., 2022) collects US political news articles from AllSides website. The articles that discuss the same event and present different ideological views are grouped together as a cluster. Each cluster contains three articles, and each article comes from liberal, center, conservative media sources respectively. The dataset also provides an expert written summary for each cluster of articles. We follow the dataset splitting released by Lee et al. (2022), which results in 2452 / 307 / 307 news clusters allocated to the train, valid, test sets.

5.2 Experimental Settings

To validate our approach, we use two types of language models as the foundation model for summarization in the experiments: a decoder-only model and a encoder-decoder model. For the decoder-only model, we choose the open-source large language

	Content Evaluation					Bias Evaluation			
	Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum	BLEU-2	polarization	p-arousal	n-arousal	sum-arousal
Baselines									
LexRank	42.24	18.16	26.61	36.87	17.68	63.73	2.42	1.53	3.95
BART-CNN	38.22	15.73	25.52	34.25	15.26	76.67	2.02	1.14	3.16
BART-Multi	39.60	16.26	24.57	35.05	17.60	54.47	3.46	1.64	5.10
Pegasus-CNN	38.17	15.61	25.18	31.17	13.69	74.45	1.98	1.15	3.12
Pegasus-Multi	35.33	13.27	21.25	31.14	14.79	59.15	4.83	2.27	7.10
GPT-3.5	42.01	16.25	26.13	37.27	18.77	77.25	3.40	2.13	5.52
GPT-3.5 + one-shot	41.95	16.77	28.13	37.39	18.33	44.29	2.52	1.57	4.08
GPT-3.5 + graph	43.20	18.06	30.56	38.08	18.99	32.37	2.03	1.59	3.62
GPT-4	42.36	16.49	26.30	37.31	19.04	75.86	3.37	1.97	5.34
GPT-4 + graph	42.61	18.67	30.82	38.18	19.09	31.77	2.11	1.49	3.60
NeuS	39.09	18.93	29.74	35.35	16.21	38.51	1.69	0.83	2.53
Bang et al. (2023)	-	-	-	-	-	-	1.57	0.91	2.48
LED	40.30	18.63	30.24	36.26	17.30	31.97	1.59	0.86	2.45
+ textual graph	41.84	19.46	31.18	37.30	18.23	29.77	1.29	0.83	2.12
+ graph prompt	42.06	19.96	31.74	37.56	18.23	29.84	1.33	0.84	2.17
+ both (full model)	42.96	20.66	32.74	38.56	19.09	28.14	**1.26**	**0.71**	**1.97**
Llama-2	42.26	19.25	30.88	37.75	19.15	30.30	1.80	1.02	2.81
+ textual graph	43.98	20.52	32.57	39.30	20.18	28.22	1.59	0.94	2.53
+ graph prompt	44.44	21.01	32.95	39.97	20.42	28.01	1.50	1.01	2.50
+ both (full model)	**45.14**	**22.30**	**34.02**	**40.74**	**21.89**	**27.89**	1.55	0.90	2.46

Table 1: Automatic Evaluation results of neutralized summarization on NeuS dataset. We calculate the Rouge and BLEU scores to evaluate content preservation; and we calculate the polarization score and the arousal scores to evaluate content-level informational bias and lexical-level bias respectively. The summarizer with better performance should attain higher Rouge and BLEU scores, but lower polarization score and arousal scores.

model LLama-2 and use the version of llama-2-7b-chat-hf (Touvron et al., 2023). For the encoder-decoder model, considering the input text is typically long, we choose the longformer encoder-decoder (LED) model and use the version of led-large-16384 (Beltagy et al., 2020).

The models take the instruction prompt and a cluster of three articles as input, and generates a summary as output. The instruction prompt provided to the Llama-2 and LED models is detailed in Appendix B. The maximum input length is set as 2048, maximum output length is 512, number of epochs is 5, gradient accumulation step is 16, weight decay is 1e-2, learning rate for Llama-2 is 1e-4 and learning rate for LED is 1e-5. The Llama-2 model is trained with LoRA (Hu et al., 2021), with the rank 8, alpha 16 and dropout 0.05.

5.3 Baselines

We implemented baselines that are mentioned in Lee et al. (2022) for comparison. Furthermore, we also establish several GPT-based baselines.

LexRank (Erkan and Radev, 2004) is an unsupervised model that selects sentences based on graph centrality and generates extractive summaries.

BART-CNN (Lewis et al., 2020) is a news summarization model that fine tunes BART-large on the CNN Daily Mail dataset (Nallapati et al., 2016).

BART-Multi (Lewis et al., 2020) is a multi-news summarization model that fine tunes BART-large on the Multi-News dataset (Fabbri et al., 2019).

Pegasus-CNN (Zhang et al., 2020) is a news summarization model that fine tunes Pegasus-large on the CNN Daily Mail dataset (Nallapati et al., 2016).

Pegasus-Multi (Zhang et al., 2020) is a multi-news summarization model that fine tunes Pegasus-large on the Multi-News dataset (Fabbri et al., 2019).

NeuS (Lee et al., 2022) develops an abstractive summarization method that learns to generate summary in a hierarchical order from title to article.

Bang et al. (2023) designs a polarity minimization loss function to reduce framing bias.

GPT-3.5 is a large language model that generates abstractive summaries via prompting. We use gpt-3.5-turbo version and prompt is in Appendix C.

GPT-3.5 + one-shot provides one example of three articles and their neutralized summary as a demonstration into the gpt-3.5-turbo model.

GPT-3.5 + graph guides the gpt-3.5-turbo model to firstly reason the event relation graph of the given articles, and then generate the summary through a chain-of-thought process (Wei et al., 2023).

GPT-4 is another large language model that automatically generates abstractive summaries. We use the gpt-4 version to create the summaries.

	Bias Evaluation		Content Evaluation		Language Evaluation	
	Lexical Bias	Informational Bias	Non-Hallucination	Recovery	Fluency	Coherency
NeuS	85.90	88.46	71.79	68.42	61.54	76.92
GPT-4	80.77	79.49	**89.74**	**97.43**	97.43	**97.43**
LED	81.58	78.20	73.68	73.68	84.21	89.74
LED + graph	**92.10**	85.90	85.00	78.95	89.47	92.31
Llama-2	83.33	84.61	68.42	74.36	**100.00**	94.87
Llama-2 + graph	91.02	**89.74**	84.21	87.18	**100.00**	**97.43**

Table 2: Human Evaluation results of neutralized summarization on NeuS dataset. The bias evaluation includes both lexical bias and informational bias. The higher scores for the six metrics represent better performance. The row "+ graph" means incorporating the multi-doc event relation graph.

	Content Evaluation					Bias Evaluation			
	Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum	BLEU-2	polarization	p-arousal	n-arousal	sum-arousal
Llama-2	42.26	19.25	30.88	37.75	19.15	30.30	1.80	1.02	2.81
+ event moral	43.82	20.65	32.48	39.42	20.11	29.05	1.58	0.93	2.51
+ in-doc relations	44.74	21.31	33.11	40.22	21.01	28.57	1.68	1.00	2.68
+ cross-doc coreference	44.53	20.78	32.80	39.53	20.72	28.16	1.63	0.97	2.60
+ all (full model)	**45.14**	**22.30**	**34.02**	**40.74**	**21.89**	**27.89**	**1.55**	**0.90**	**2.46**

Table 3: The ablation study of different components in the multi-document event relation graph. The summarizer with better performance should attain higher Rouge and BLEU scores, lower polarization score and arousal scores.

GPT-4 + graph incorporates the graph into gpt-4 model through a chain-of-thought process.

5.4 Automatic Evaluation

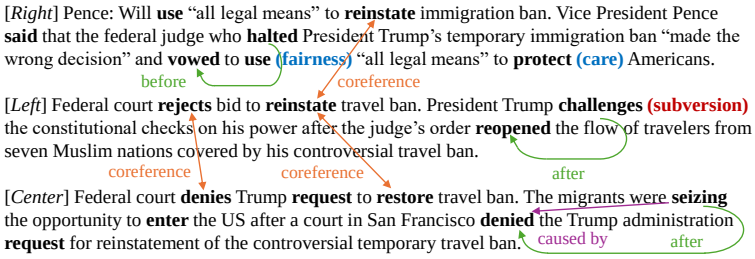
The automatic evaluation aims to evaluate the summarization models from the perspective of content preservation and bias mitigation. To evaluate content preservation, we calculate Rouge scores (Lin, 2004) and BLEU score (Papineni et al., 2002) between model generated summaries and human written reference. As bias can be induced by both emotional language and biased information included in the content (Fan et al., 2019), to evaluate bias mitigation, we evaluate both lexical-level bias and content-level informational bias, by calculating the arousal scores and polarization score respectively (calculation detailed in Appendix D). The arousal scores designed by Lee et al. (2022) include three metrics: positive-arousal, negative-arousal, and sum-arousal. A better summarizer is expected to attain higher Rouge and BLEU scores, as well as lower polarization score and arousal scores. The automatic evaluation results are shown in Table 1.

The results demonstrate that incorporating the multi-document event relation graph effectively mitigates both lexical and informational bias, and meanwhile improving content preservation for both Llama-2 and LED models. Compared to their baseline models without the graph, our approach successfully alleviates media bias and reduces the polarization score and arousal scores. Besides, our approach notably improves content preservation and yields higher Rouge and BLEU scores. In addition, incorporating the multi-document event relation graph effectively improves content preservation as well as alleviates media bias for commercial GPT models as well, reducing polarization and arousal scores and meanwhile yielding higher Rouge and BLEU scores. Moreover, our approach based on the multi-document event relation graph outperforms previous methods that lack bias information. Overall, the finetuned Llama-2 + graph model (the very last row) achieves the highest Rouge scores, BLEU score, and the lowest polarization score, and the finetuned LED + graph model (the last row of LED models) exhibits the lowest arousal scores.

5.5 Human Evaluation

The human evaluation aims to evaluate the generated summaries from three perspectives: bias mitigation, content semantics, and language quality. Specifically, we design six questions regarding lexical bias (usage of emotional or biased words), informational bias (presentation of biased information that reflects ideological leanings), content non-hallucination, content recovery, language fluency, and language coherency. After collecting the evaluation scores from human annotators for the six questions, we normalize the metrics into the range of zero to one. For all the six metrics, a higher score is better. The details about human evaluation are presented in Appendix E. The results of human evaluation are presented in Table 2.

The human evaluation validates the consistent



[*llama baseline*] A federal judge in Seattle temporarily **blocked** President Trump's executive order that **barred** immigration from seven countries. The judge said the order was unconstitutional and Trump **challenged** the constitutional check. *(bleu 7.35, arousal 3.31)*

[*llama with multi-doc event relation graph*] A federal court has **blocked** the travel ban that was put in place by President Trump. The court has also **denied** the Trump Administration's **request** to have the ban **reinstated**. *(bleu 15.59, arousal 1.96)*

Figure 3: A qualitative analysis of the generated text before and after incorporating the multi-document event relation graph. The multi-document event relation graph effectively mitigates framing bias for neutralized summarization.

observations with automatic evaluation: the approach based on the multi-document event relation graph effectively mitigates both lexical bias and informational bias. This demonstrates that the multi-document event relation graph informs the LLMs of bias distribution, and guides the LLMs to utilize less biased words as well as present less biased information. Besides, the multi-document event relation graph also improves content quality, by reducing the hallucination and enhancing content recovery. The multi-document event relation graph explicitly show the reported events and represents the event-level content structures within and across documents, thereby assists in reducing hallucinations and recovering salient events.

Different from automatic evaluation, human evaluation indicates that GPT-4 model shows the best content quality among the evaluated models, with the least hallucination and best content recovery. One explanation is that Rouge and BLEU scores are calculated based on human written references, while human evaluation measures content non-hallucination and content recovery rate with respect to the input articles. Both automatic evaluation and human evaluation complement each other to achieve a comprehensive assessment of summary quality. Despite its strong performance in content quality, the GPT-4 model still suffers from a higher level of lexical bias and informational bias. This shows that the current powerful LLMs can still carry ideological bias in the generated content and mitigating ideological bias is necessary.

5.6 Ablation Study

The ablation study of the two designed strategies is shown in Table 1. The results demonstrate that both textualized graph and graph prompt play a necessary role in mitigating media bias and improving content preservation. The textual graph and graph prompt complement each other: the textual graph augments the instruction prompt with graph information, while the graph prompt enables direct learning from the graph embedding. Incorporating the two strategies together achieves the best performance for both Llama-2 and LED models.

The ablation study of different components in the multi-document event relation graph is presented in Table 3. The results show that all the components in the graph are critical in mitigating media bias. The different elements in the graph carry complementary bias information: event-level moral opinions inform opinionated moral bias, within-doc event relations reflect different narrative framings of each article, and cross-doc event coreference relation highlights event selection bias across various articles. All these components are useful in constructing a unified and cohesive content structure. Integrating them together as a whole graph yields the best performance in terms of both content preservation and bias mitigation.

5.7 Qualitative Analysis

Figure 3 shows a qualitative analysis of the generated summaries before and after incorporating the multi-document event relation graph into Llama-2 model (An example of the LED model is shown in Appendix F). The generated summary from the Llama-2 baseline (upper text) does not mention the core event *denies the reinstatement*, but includes biased information *challenged the constitutional check* from the liberal article that presents the liberal ideological bias, and also has hallucination *a federal judge from Seattle* which contradicts with the input *a court in San Francisco*. After incorporating the multi-document event relation graph, the generated summary (lower text) brings the mutually reported events *denied the reinstatement* back, successfully excludes the biased information, and eliminates the hallucination. This shows the effectiveness of multi-document event relation graph

in improving content preservation and mitigating media bias.

6 Conclusion

This paper aims to generate a neutralized summary given multiple articles with differing ideological bias as input, thereby mitigating their framing bias. Motivated by the critical roles of events and event relations in detecting media framing bias, we propose to build a multi-document event relation graph to inform LLMs of bias distribution. We further design two strategies to incorporate the multi-document event relation graph into LLMs for guiding the multi-document summarization process, which include graph textualization and graph prompt tuning. Both automatic evaluation and human evaluation demonstrate the effectiveness of our approach in mitigating media bias and meanwhile improving content preservation.

Limitations

Our paper proposes to construct a multi-document event relation graph to guide the neutralized summarization process. The performance of this graph is not perfect and may still make errors in extracting event relations. We observe that the current event relation graph has the ability to extract event relations for most easy cases with explicit discourse connectives or language cues. However, it may make mistakes in recognizing hard cases that state event relations in an implicit way. To further improve the performance of media bias mitigation, it is necessary to enhance the extraction of implicit event relations. Therefore, improving the construction of multi-document event relation graph becomes necessary and serves as the future work.

Ethical Considerations

This paper develops methodology to mitigate media bias. The media framing bias is a type of unwanted bias, which has harmful impact on both individuals and the society, such as misleading readers, intensifying societal polarization, and undermining democratic values. The goal of this paper is to mitigate the unwanted media framing bias and enhance unbiased information access. The examples in this paper are only used for research purpose, and do not represent any political leaning of the authors. The release of code and model should be leveraged to address and reduce media bias, serving a broader social good.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback and input. We gratefully acknowledge support from National Science Foundation via the awards IIS-2127746 and IIS-1942918. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High-Performance Research Computing.

References

Kent Asp. 2007. Fairness, informativeness and scrutiny: The role of news media in democracy. *Nordicom Review*, 28.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, page 79–85, USA. Association for Computational Linguistics.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 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)*, pages 4982–4991.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765*.

Yejin Bang, Nayeon Lee, and Pascale Fung. 2023. Mitigating framing bias with polarity minimization loss. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11100–11110, Singapore. Association for Computational Linguistics.

David P Baron. 2005. Competing for the public through the news media. *Journal of Economics & Management Strategy*, 14(2):339–376.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271.

Jie Cai and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 28–36, Tokyo, Japan. Association for Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Hong Chen, Raphael Shu, Hiroya Takamura, and Hideki Nakayama. 2021. GraphPlan: Story generation by planning with event graph. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 377–386, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al-Khatib, Benno Stein, and Henning Wachsmuth. 2020. Detecting media bias in news articles using gaussian bias distributions. *arXiv preprint arXiv:2010.10649*.

Peng Cui and Le Hu. 2021. Topic-guided abstractive multi-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1463–1472, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Paul D'Angelo. 2017. Framing: media frames. *The international encyclopedia of media effects*, pages 1–10.

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

Amir Emami, Dianne HB Welsh, Veland Ramadani, and Ali Davari. 2020. The impact of judgment and framing on entrepreneurs' decision-making. *Journal of Small Business & Entrepreneurship*, 32(1):79–100.

Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.

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

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

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 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)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Matthew Gentzkow and Jesse M Shapiro. 2006. Media bias and reputation. *Journal of political Economy*, 114(2):280–316.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Tim Groeling. 2013. Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, 16:129–151.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 434–444, Hong Kong, China. Association for Computational Linguistics.

Claudia Hillebrand. 2019. The role of news media in intelligence oversight. In *Secret Intelligence*, pages 396–412. Routledge.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *Preprint*, arXiv:1508.01991.

Xiaomeng Jin, Manling Li, and Heng Ji. 2022. Event schema induction with double graph autoencoders. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2013–2025, Seattle, United States. Association for Computational Linguistics.

Jim A Kuypers. 2002. *Press bias and politics: How the media frame controversial issues*. Bloomsbury Publishing USA.

Tuan Lai, Heng Ji, Trung Bui, Quan Hung Tran, Franck Dernoncourt, and Walter Chang. 2021. A context-dependent gated module for incorporating symbolic semantics into event coreference resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages

3491–3499, Online. Association for Computational Linguistics.

Viet Lai, Hieu Man, Linh Ngo, Franck Dernoncourt, and Thien Nguyen. 2022. Multilingual SubEvent relation extraction: A novel dataset and structure induction method. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5559–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. NeuS: Neutral multi-news summarization for mitigating framing bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.

Yuanyuan Lei and Houwei Cao. 2023. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing*, 14(4):2954–2969.

Yuanyuan Lei and Ruihong Huang. 2022. Few-shot (dis)agreement identification in online discussions with regularized and augmented meta-learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5581–5593, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2023a. Discourse structures guided fine-grained propaganda identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 331–342, Singapore. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2023b. Identifying conspiracy theories news based on event relation graph. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9811–9822, Singapore. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2024a. Boosting logical fallacy reasoning in LLMs via logical structure tree. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13157–13173, Miami, Florida, USA. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2024b. Sentence-level media bias analysis with event relation graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5225–5238, Mexico City, Mexico. Association for Computational Linguistics.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuanyuan Lei, Md Messal Monem Miah, Ayesha Qamar, Sai Ramana Reddy, Jonathan Tong, Haotian Xu, and Ruihong Huang. 2024a. Emona: Event-level moral opinions in news articles. *Preprint*, arXiv:2404.01715.

Yuanyuan Lei, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Ruihong Huang, and Dong Yu. 2024b. Polarity calibration for opinion summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5211–5224, Mexico City, Mexico. Association for Computational Linguistics.

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

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020. Connecting the dots: Event graph schema induction with path language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 684–695, Online. Association for Computational Linguistics.

S Robert Lichter. 2017. Theories of media bias. *The Oxford handbook of political communication*, pages 403–416.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022a. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Yujian Liu, Xinliang Zhang, Kaijian Zou, Ruihong Huang, Nicholas Beauchamp, and Lu Wang. 2023. All things considered: Detecting partisan events from

news media with cross-article comparison. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15472–15488, Singapore. Association for Computational Linguistics.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022b. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Hieu Man, Minh Nguyen, and Thien Nguyen. 2022. Event causality identification via generation of important context words. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 323–330, Seattle, Washington. Association for Computational Linguistics.

Yuning Mao, Yanru Qu, Yiqing Xie, Xiang Ren, and Jiawei Han. 2020. Multi-document summarization with maximal marginal relevance-guided reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1737–1751, Online. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184.

Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R Corman, and Huan Liu. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1(2):1–18.

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

Navakanth Reddy Naredla and Festus Fatai Adedoyin. 2022. Detection of hyperpartisan news articles using natural language processing technique. *International Journal of Information Management Data Insights*, 2(1):100064.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Richard Yuanzhe Pang, Adam Lelkes, Vinh Tran, and Cong Yu. 2021. AgreeSum: Agreement-oriented multi-document summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3377–3391, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. 2021. Efficiently summarizing text and graph encodings of multi-document clusters. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4768–4779, Online. Association for Computational Linguistics.

Markus Prior. 2013. Media and political polarization. *Annual review of political science*, 16:101–127.

M. RECASENS and E. HOVY. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Eitan Sapiro-Gheiler. 2019. Examining political trustworthiness through text-based measures of ideology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 10029–10030.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. A hierarchical encoding-decoding scheme for abstractive multi-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5872–5887, Singapore. Association for Computational Linguistics.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2021. Extracting event temporal relations via hyperbolic geometry. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8065–8077, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James W Tankard Jr. 2001. The empirical approach to the study of media framing. In *Framing public life*, pages 111–121. Routledge.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Esther van den Berg and Katja Markert. 2020. Context in informational bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6315–6326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Francielle Vargas, Kokil Jaidka, Thiago Pardo, and Fabrício Benevenuto. 2023. Predicting sentence-level factuality of news and bias of media outlets. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1197–1206, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhailong Wang, Yi Fung, Piyush Mishra, Qing Lyu, Dídac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 133–143, Online. Association for Computational Linguistics.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

Wenlin Yao, Zeyu Dai, Maitreyi Ramaswamy, Bonan Min, and Ruihong Huang. 2020. Weakly supervised subevent knowledge acquisition. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pages 5345–5356.

Ruifeng Yuan, Zili Wang, and Wenjie Li. 2021. Event graph based sentence fusion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4075–4084, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2020. Event coreference resolution with their paraphrases and argument-aware embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *Preprint*, arXiv:1912.08777.

Xinliang Frederick Zhang, Winston Wu, Nick Beauchamp, and Lu Wang. 2024. Moka: Moral knowledge augmentation for moral event extraction. *Preprint*, arXiv:2311.09733.

Xiyang Zhang, Muhao Chen, and Jonathan May. 2021. Salience-aware event chain modeling for narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1418–1428, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hao Zhou, Weidong Ren, Gongshen Liu, Bo Su, and Wei Lu. 2021. Entity-aware abstractive multi-document summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 351–362, Online. Association for Computational Linguistics.

A Multi-document Event Relation Graph

A.1 Implementation Details

An event identification model is trained on the MAVEN dataset which annotated event mentions for general-domain documents (Wang et al., 2020). Considering the news articles are usually long, we use the Longformer (Beltagy et al., 2020) language model to encode the article and build a binary classification head on top of the word embeddings to predict whether the word triggers an event or not.

The event-level moral opinions classifier is trained based on the EMONA dataset which annotated moral opinions towards each event in news articles (Lei et al., 2024a). Following Lei et al. (2024a), we use the Longformer (Beltagy et al., 2020) to encode the entire article and add an extra layer of Bi-LSTM (Huang et al., 2015) on top to capture the contextual information. Then we build a 11-class classification head on top of the event words embeddings to predict the moral label. The 11 labels include ten moral foundations (care, harm, fairness, cheating, loyalty, betrayal, authority, subversion, purity, degradation) and the non-moral label.

The four event relation extractors are trained on the general-domain MAVEN-ERE dataset (Wang et al., 2022). We follow previous work (Han et al., 2019; Yao et al., 2020) to form the training event pairs in natural textual order, meaning the former event in the pair is the precedent event mentioned in text. For the temporal relations, the dataset also annotates the time expressions such as date or time. Considering our event relation graph focuses on events, we only retain annotations between events. We further process the *before* annotation in this way: keep the *before* label if the annotated event pairs aligns with the natural textual order, otherwise assign *after* label to the reverse pair. The *simultaneous*, *overlap*, *begins-on*, *ends-on*, *contains* annotations are grouped into *overlap* category in our graph. For the causal relations, we keep the *cause* label if the natural textual order is followed, or assign the *caused by* label if not. For the *subevent* relations, we maintain the *contain* label if the natural textual order is followed, and assign *contained by* label otherwise. The Longformer (Beltagy et al., 2020) is used as the foundation language model, and the event pair embedding is the concatenation of two event words embeddings. Since the four event relations interact with each other to form a cohesive narrative structure, we adopt the joint

learning framework (Wang et al., 2022) to train these relations collaboratively.

A.2 Evaluation Performance

Table 4 presents the performance of event identification. Table 5 shows the performance of event-level moral opinions classification, where we use macro precision, recall, and F1 score as evaluation metrics. Table 7 shows the performance of event coreference resolution. Following the previous work (Cai and Strube, 2010), MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_e$ (Luo, 2005), and BLANC (RECASENS and HOVY, 2011) are used as evaluation metrics. The performances of other components in the event relation graph, including temporal, causal, and subevent relation classification are summarized in Table 6. The standard macro-average precision, recall, and F1 score are reported.

	Precision	Recall	F1
Event Identifier	87.31	91.81	89.40

Table 4: Performance of event identification. Macro precision, recall, and F1 are reported.

	Precision	Recall	F1
Moral Classifier	45.75	38.94	41.13

Table 5: Performance of event-level moral opinions classification. Macro precision, recall, and F1 are reported.

	Precision	Recall	F1
Temporal	48.45	46.43	47.04
Causal	58.48	54.02	56.01
Subevent	53.37	42.90	46.21

Table 6: Performance of temporal, causal, and subevent relation tasks in the event relation graph. Macro precision, recall, and F1 are reported.

A.3 Statistical Analysis

Table 8 presents the statistics of multi-doc event relation graph on NeuS dataset. On average, there are 25.59 events, 2.81 moral events, 1.82 coreference relation, 38.64 temporal relation, 6.27 causal relation, 1.24 subevent relation, and 3.42 cross-doc coreference relation within one graph.

B Instruction Prompt for Summarization

The instruction prompt provided into Llama-2 and LED baseline models is: "Please summarize the given text. Text: <article 1> /s <article 2> /s <article 3>. Summary:"

The instruction prompt to incorporate the textualized graph into Llama-2 and LED models is: "The task is summarizing the given text. The events and event relations are important for summarization. An event is an occurrence or action reported in the text. The moral attribute of event represents the event is objective or contains subjective moral evaluation. The following table presents the events in the text, and the columns are event id, event word, and event moral attribute: <event table>. There are four types of event relations: coreference, temporal, causal, and subevent relations. Coreference relation represents two events designate the same occurrence. Temporal relation represents the chronological order between events, such as before, after, and overlap. Causal relation represents the causality between events. Subevent relation represents the containment relation from a parent event to a child event. The following table presents the event relations in the text, and the columns are source event id, source event word, relation between source event and target event, target event id, target event word: <event relation table>. Please summarize the given text. Text: <article 1> /s <article 2> /s <article 3>. Summary:"

C Prompt for GPT-based baselines

The prompt provided into gpt-3.5-turbo abd gpt-4 baselines is: "Please summarize the given text. Text: <article 1> /s <article 2> /s <article 3>. Summary:"

The prompt provided into gpt-3.5-turbo + one-shot baseline is: "Please summarize the given text. Please mimic the output style in the following example. Example: <example article 1> /s <example article 2> /s <example article 3>. Summary: <example summary>. Text: <article 1> /s <article 2> /s <article 3>. Summary:"

The prompt provided into gpt-3.5-turbo + graph baseline is: "Please summarize the given text. Let's think step by step. Firstly, explain the events reported in each article and the relations between events. Events refer to an occurrence or action reported in the sentence. There are four types of event relations: coreference, temporal, causal, and subevent relations. Coreference relation represents two events designate the same occurrence. Temporal relation represents the chronological order between events, such as before, after, and overlap.

MUC			B^3			$CEAF_e$			BLANC		
Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
76.34	83.10	79.57	97.07	98.32	97.69	97.79	97.00	97.39	83.69	92.43	87.54

Table 7: Performance of event coreference resolution in the event relation graph

events	moral events	event pairs	coreference	temporal	causal	subevent	cross-doc coref
25.59	2.81	118.53	1.82	38.64	6.27	1.24	3.42

Table 8: The statistics of multi-document event relation graph on NeuS dataset. The average number of events, moral events, event pairs, four in-doc event relations, and cross-doc event coreference relation in a graph are shown.

Causal relation represents the causality between events. Subevent relation represents the containment relation from a parent event to a child event. Secondly, generate the summary. Please mimic the output style in the following example. Example: <example article 1> /s <example article 2> /s <example article 3>. Output: Firstly, explain the events reported in each article and the relations between events. <explanation of events and event relations in the example articles> Secondly, generate the summary. Summary: <example summary>. Text: <article 1> /s <article 2> /s <article 3>. Output:"

D Automatic Evaluation

The automatic evaluation aims to evaluate the summarization models from the perspective of content preservation and bias mitigation.

For the content evaluation, we use Rouge scores (Rouge-1, Rouge-2, Rouge-L, Rouge-Lsum scores) and BLEU score (cumulative BLEU-2 score) between model generated summaries and human written reference as metrics (Lin, 2004; Papineni et al., 2002). The Rouge scores are implemented by using the rouge_scorer.RougeScorer function from the rouge_score package. The cumulative BLEU-2 score is implemented by using the sentence_bleu function from the nltk.translate.bleu_score package, with a weight of (0.5, 0.5, 0, 0).

For the bias evaluation, we evaluate both lexical-level bias and content-level informational bias, by using the arousal scores and polarization score respectively. This is because bias can be induced by both emotional language and biased information within content. The arousal scores are designed by Lee et al. (2022) and evaluate lexical-level emotion. Specifically, the arousal scores are computed based on the Valence-Arousal-Dominance emotion lexicon (Mohammad, 2018) and include three metrics: p-arousal (arousal scores of positive words in the summary), n-arousal (arousal scores of negative words), and sum-arousal (summation of positive and negative arousal). On the other hand, the polarization score evaluates content-level informational bias. To calculate polarization score, an ideology classifier is trained on the AllSides dataset (Baly et al., 2020), where a classification head is built on top of the BERT language model to classify article-level ideology into liberal, center, and conservative. The macro F1 score of this ideology classifier is 80.19, which proves its validity. The polarization score is computed as the predicted probability of the polarized class, i.e. the sum of predicted probability of liberal and conservative classes. The polarization score ranges from 0 to 1, with 1 indicating full ideological polarization (liberal or conservative) and 0 indicating no polarization (neither liberal nor conservative).

E Human Evaluation

The human evaluation aims to evaluate the generated summaries from three perspectives: bias mitigation, content semantics, and language quality. Considering bias can be induced by both lexical usage and information selection (Fan et al., 2019), we evaluate both types of lexical bias (usage of emotional or biased words) and informational bias (presentation of biased information that reflects ideological leanings). Specifically, we design six questions regarding lexical bias, informational bias, content non-hallucination, content recovery, language fluency, and language coherency. The detailed questions are:

1. Question (lexical **bias**): Does the text use biased or emotional words? Choose 2, 1, 0. Score Explanation: 2 - uses objective and neutral words, 1 - uses some emotional words but acceptable, 0 - uses biased and emotional words.

2. Question (informational **bias**): Does the text

	generated summary
golden summary	Travel Ban Reinstatement Rejected. A federal court rejected the reinstatement of the travel ban. The migrants from the seven listed countries are allowed to enter the US.
LED	Vice President Mike Pence said Sunday that the federal judge who halted President Trump's temporary immigration ban "made the wrong decision" and vowed to use "all legal means at our disposal" to protect Americans.
LED + graph	Appeals Court Denies Request to Immediately Reinstate Travel Ban. A federal appeals court denied the Trump administration's request to immediately reinstate the travel ban.
Llama-2	A federal judge in Seattle temporarily blocked President Trump's executive order that barred immigration from seven countries. The judge said the order was unconstitutional and Trump challenged the constitutional check.
Llama-2 + graph	A federal court has blocked the travel ban that was put in place by President Trump. The court has also denied the Trump Administration's request to have the ban reinstated.

Table 9: A qualitative example of the generated summaries before and after incorporating the multi-document event relation graph into LED and Llama-2 models.

contain biased or polarized information to showcase ideology leaning? Choose 2, 1, 0. Score Explanation: 2 - does not showcase ideology leaning, 1 - conveys some ideology bias but acceptable, 0 - contains biased information to showcase ideology leaning.

3. Question (**content** non-hallucination): Does the text hallucinate compared to input articles? Choose 1 or 0. Score Explanation: 1 - does not hallucinate and aligns with the facts in input articles, 0 - hallucinates some content that is not included in the input articles.

4. Question (**content** recovery): Does the text recover the important content from input articles? Choose 1 or 0. Score Explanation: 1 - successfully recover important content from input articles, 0 - misses some important content in input articles.

5. Question (**language** fluency): Is the text fluent and grammarly correct? Choose 1 or 0. Score Explanation: 1 - fluent and grammarly correct, 0 - not fluent and has grammar errors.

6. Question (**language** coherency): Is the text coherent with natural logic flow? Choose 1 or 0. Score Explanation: 1 - coherent and logic flow is natural, 0 - not coherent and logic flow is not natural.

There are two human annotators participated in the evaluation, both are graduate students who are familiar with natural language processing and media bias research. Their reported political leanings are center. One annotator is native English speaker and the other is proficient in English. Both the annotators were paid. The Cohen's kappa inter-annotator agreement rate is 0.67. The randomly sampled 50 clusters of articles from the test set were annotated. We select the NeuS and GPT-4 baselines, LED, LED + graph (the last row of LED models in Table 1), Llama-2, Llama-2 + graph (the very last row in Table 1) for evaluation. To avoid the leakage of model information, different models are randomly shuffled and the name of models are omitted. After collecting the evaluation scores from human annotators for the six questions, we normalize the metrics into the range of zero to one. For all the six metrics, a higher score is better.

F Qualitative Analysis

Table 9 presents an example of the generated summaries from LED and LED + graph models, as well as the summaries from Llama-2 and Llama-2 + graph models. The multi-document event relation graph alleviates the ideological leaning in the summaries for both LED and Llama-2 models.