LLM-based Ideological Discourse Analysis in News Events using Contextualized Semantic Discourse Structure

Anonymous ACL submission

Abstract

Analyzing ideological discourse even in the age of LLMs remains a challenge, as these models often struggle to capture the key elements that shape real-world narratives. Specifically, LLMs fail to focus on characteristic elements driving dominant discourses and lack the ability to integrate contextual information required for understanding abstract ideological views. To address these limitations, we propose a framework motivated by the theory of ideological discourse analysis that constructs a vocabulary of repeating "talking points" from a corpus of news articles related to real-world events. Our framework contextualizes these talking points by analyzing the interaction between entities, their roles, and media frames to generate ideology-specific viewpoints (or partisan perspectives). We evaluate our framework's ability to generate these perspectives through automated tasks— ideology and partisan classification tasks, supplemented by human validation. Additionally, we also demonstrate straightforward applicability of our framework in creating visual event snapshots, providing a new way of interpreting real-world events. We release resulting dataset and model to the community to support further research.

1 Introduction

One of the signs of the growing social and political polarization is the formation of insulated information bubbles (Gentzkow and Shapiro, 2011; Quattrociocchi et al., 2016; Dubois and Blank, 2018; Garimella et al., 2018), in which news media discourse is shaped around ideological lines, often intended to shape the readers' views. Understanding this phenomenon better, to the extent we can examine the degree to which members of the two communities hold opposite accounts of reality, requires computational methods that can compare the narratives of both sides and identify the points in which their accounts converge and diverge (example shown in Figure 1).

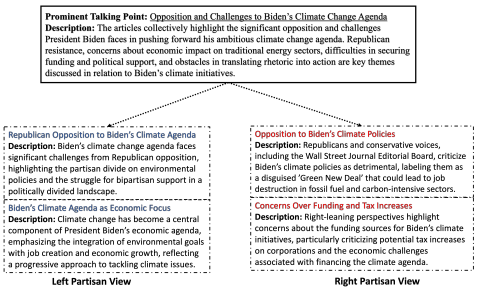


Figure 1: An example of a repeating talking point with its partisan viewpoints for an event about *Climate Change*.

Past work analyzing political discourse typically focused on discrete aspects, such as stance and bias detection works (Liu et al., 2022; Luo et al., 2020; Kiesel et al., 2019; Li and Goldwasser, 2019), political news framing (Mendelsohn et al., 2021; Field et al., 2018; Card et al., 2015b), sentiment toward relevant entities (Park et al., 2021; Rashkin et al., 2016), which while relevant, fall short of providing the comprehensive view needed. The rise of Large Language Models (LLM) has a transformative potential for enabling complex narrative analysis connecting these dimensions and explaining their relationship. However, realizing it is not straightforward, as demonstrated by several recent works analyzing political texts, either as a straightforward stance prediction task (Ziems et al., 2024), or mapping political positions to specific stances on policy issues (Santurkar et al., 2023).

To address these challenges, we suggest a structured approach, modeling the interactions between different narrative elements. Drawing inspiration from foundational political discourse analysis (Van Dijk, 1996; Van Dijk et al., 1997), we center our study around the notion of *talking point*, a semantic discourse structure that captures a specific aspect of news event through a short summary and a set of properties – the lens through which it is discussed, using media-frames (Boydstun et al., 2014), relevant entities, their roles and attitudes towards them (Khanehzar et al., 2021; Roy et al.,

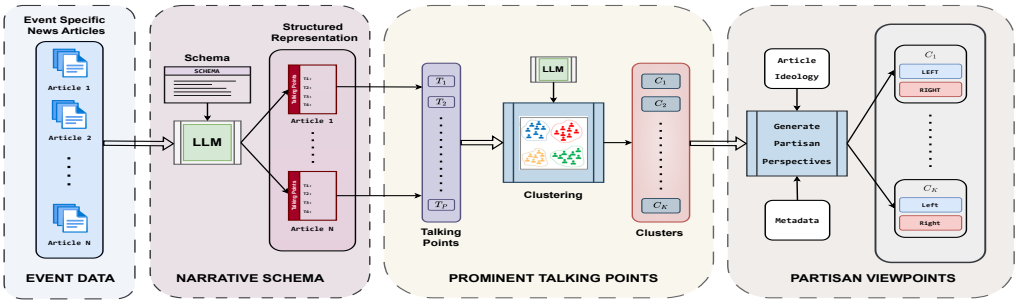


Figure 2: Provides an overview of our framework. The event-specific news articles are structured according to our narrative schema, which helps identify recurring talking point vocabulary (*prominent talking points*). These prominent points are then used to derive partisan perspectives.

2021). We analyze political discourse by creating a unified vocabulary of repeating talking points, and comparing their differences across the political sides, using these properties. This analysis allows us to identify *agenda-setting* attempts (McCombs and Shaw, 1972; Scheufele and Tewksbury, 2007) by looking at frequent talking points overwhelmingly discussed by one side, as well as *areas of consensus and polarization*, based on talking points frequently discussed by both sides, in a similar or contrasting way. Figure 4 exemplifies this analysis for Covid-19.

To accomplish that, we suggest a pipeline approach (described in Figure 2). First, we use an LLM for extracting information from each article using a structured schema. Second, we create a talking point vocabulary by clustering the extracted schema information and using an LLM to improve the clusters by removing inconsistent and redundant clusters. The goal of this process is to identify repeating topics that can be potentially shared by both political parties. Finally, we use an LLM to reason about the partisan view of each recurring point, capturing the differences in framing and attitudes toward entities expressed by each side.

This process generates resources validated by both human and automated techniques, beneficial for researchers. First, we provide a dataset of narratives, extracted from 6,141 articles on 24 events related to 4 contested political issues, sourced from 126 outlets coded for bias[1]. Additionally, we identify 389 higher-level talking points that reflect both left- and right-perspectives on these events, which we refer to as ideology-specific viewpoints (or partisan viewpoints).

Note that unlike manual annotation based approach that are costly and as a result focus on specific topics, we suggest an automated method, broadly applicable across a very wide range of event topics, and show that it results in high-quality information. In particular, we validate the framework's ability to generate broader partisan discourse that captures ideological nuances of political groups through an *ideology classification task* on previously unseen news articles. We also test for the degree with which generated perspectives exhibit ideology-specific behavior through a *partisan classification task*.

To summarize, our main contributions are: **(1)** We propose a new way to characterize ideological discourse in an event by constructing a vocabulary of repeating talking points that are contextualized to capture nuanced ideological positions. **(2)** We suggest an LLM-based pipeline to generate partisan viewpoints. **(3)** We perform automated and human validation of generated partisan perspectives, resulting in a novel dataset that captures partisan views over multiple events, which can be used to drive future research.

2 Related Work

Prior work on studying partisan perspectives in NLP has primarily focused on *frames*. While a contested concept, framing is commonly conceived of as a communicative structure in which the speaker highlights specific aspects of an issue to promote a political viewpoint (Goffman, 1974; Entman, 1993; Kinder, 1998; Chong and Druckman, 2007). Card et al. (2015b) proposes the Media Frame Corpus that has 15 generic media frames defined by Boydstun et al. (2014), such as economics or public opinion. In a polarized media environment, frames serve as instrumental mechanisms to promote political agendas through the selective coverage of events (informational bias) and the manipulation of their presentation (lexical bias) (Gentzkow and Shapiro, 2006; Jamieson et al., 2007; Fan et al.,

[1] https://www.allsides.com/media-bias

2019). Prior work has also explored approaches to automatically detect and mitigate framing biases. Liu et al. (2019); Akyürek et al. (2020) identify frames through news headlines, Ji and Smith (2017); Khanehzar et al. (2021) detect frames at a document level, and Lee et al. (2022); Liu et al. (2023) mitigate framing biases using multi-domain summarization and graphs. However, the formalization of frames oversimplifies the intricacies of partisan perspectives and falls short in capturing the nuance of how political agendas are deliberately conveyed in news articles. In this work, we look closer at news articles, and represent them with a predefined structure of *talking points*, carefully crafted statements that push the political messages, and cluster them to identify repeating themes, to collectively shape the different partisan perspectives. Identifying nuanced *talking points* can be thought of similar to using LLMs to generate explicit representations that helps in assessing arguments Hoyle et al. (2023). Recent work has also explored finer analysis in news articles/political biases. Lawlor and Tolley (2017) presents an entity-focused study of media news framing. Spinde et al. (2021) detects media biases at the word and sentence level, and Frermann et al. (2023) identifies and uses multi-label frames. Our work complements these by introducing a framework that allows us to establish repeating themes of talking points to unveil the partisan perspectives within an event.

3 Narrative Framework

The theory of ideological discourse analysis, as proposed by (Van Dijk, 1996; Van Dijk et al., 1997), introduces semantic discourse structures that can capture the abstract ideological positions of political groups across various social issues and contexts. In particular, (Van Dijk, 1996) suggested that ideological discourse can be predicted by examining local meanings and implications through goal and activity-descriptions, such as questions like "What are our tasks?" and "What are our roles?" etc. Additionally, (Van Dijk et al., 1997) proposed to characterize political discourse by defining semantically relevant structures known as *topics*, which are contextualized by their respective topical actors to better understand persuasive communication.

Drawing inspiration from these foundational works, we introduce a semantic discourse structure called *talking point* that enables us to characterize ideological views of political parties at an event-level granularity. Talking points serve as topic

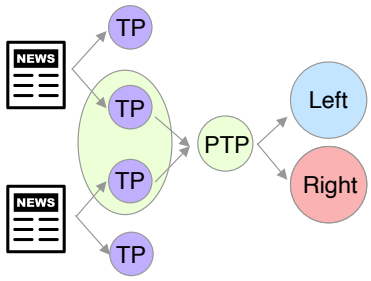


Figure 3: Our Narrative Schema: Given news articles, we extract talking points (TP), which we then cluster (green background), capturing topic relevance. Each cluster is represented by prominent talking point (PTP). Each PTP cluster captures partisan viewpoints for Left/Right political ideologies.

markers that hold semantically relevant information. These are accompanied by brief descriptions that capture key discussion elements relevant to the document (e.g., a news article). To capture nuanced ideological discourse and local meanings, we contextualize each talking point through its *metadata* - by identifying relevant entities, corresponding activities, actor/target roles pertinent to each activity, and the associated media frame.

To understand the ideological discourse around an entire news event, we focus on studying ideological views for the subset of frequently occurring talking points, referred to as *prominent talking points*. These prominent talking points repeat often to construct a narrative at an event-level granularity, and highlight specific portions of the event that are of particular interest to political parties.

The following subsections provide a detailed overview of our pipeline approach.

3.1 Narrative Schema

As the initial step in analyzing partisan discourse (illustrated in Figure 2), we represent each news article related to an event using its identified *talking points*. Given a set of n news articles $\{d_z\}_{z=1}^{n}$ relevant to an event $\mathcal{E}$, our proposed schema distills each article d_z into a set of up to four key talking points, denoted as $\{t_i\}_{i=1}^{m}$, where $m \leq 4$.

To capture local meanings and implications, we contextualize each *talking point* by extracting relevant *metadata* that can help analyze partisan differences. This metadata is constructed by first identifying the set of entities associated with each point t_i. We then capture the relationships between these entities and their influence on one another by identifying the set of activities linked to each talking point t_i, characterized by the structure *who did what to whom*. Specifically, each activity includes a brief description, an *actor* who initiates the activ-

ity, a *target* affected by the actor, and the sentiment towards the target entity, indicating whether the impact is positive or negative. Additionally, we identify the media frame associated with each activity, following the nomenclature established by Boydstun et al. (Card et al., 2015a).

To obtain the structured representations defined by our schema, we prompt an LLM to identify key talking points and the corresponding metadata using the template shown in Table 14.

3.2 Characterize Partisan Perspectives

The aim here is to analyze the political discourse surrounding an event $\mathcal{E}$ by examining how each political ideology, such as {*left*, *right*}, addresses it. We do this by identifying and focusing on prominent talking points (PTPs) that form broader narratives. Initially, each news article $\{d_z\}_{z=1}^{n} \in \mathcal{E}$ is represented using up to four talking points, $\{t_i\}_{i=1}^{m}$, $m \leq 4$. These talking points are specific to their articles, but our objective is to focus on the discourse for the entire event $\mathcal{E}$. Thus, we compile a set $\mathcal{T} = \{t_s\}_{s=1}^{p}$ by gathering all the talking points of each article in $\mathcal{E}$. We then group $\mathcal{T}$ to identify the PTP clusters and use these clusters to produce left and right viewpoints, reflecting the partisan views of the two ideologies (Figure 3). The following describes these steps.

3.2.1 Prominent Talking Points Identification

To determine the dominant discourse shaping the narrative of a political ideology within the news event $\mathcal{E}$, we start by identifying the set of *prominent talking points* that are frequently mentioned and pertinent to $\mathcal{E}$, thereby characterizing the entire set $\mathcal{T}$. This involves clustering similar talking points, where each cluster's label signifies a PTP. Clusters reflect topical similarity, and their size indicates how often the prominent point is mentioned. Algorithm 1 outlines the procedure used to obtain the prominent talking points (PTPs).

Specifically, we start by embedding each point in $\mathcal{T}$ with a dense retriever model f (Ni et al., 2021), resulting in embeddings $\mathcal{Z}$. These embeddings are then clustered using HDBSCAN (Campello et al., 2013) to obtain candidate *clusters*. For each cluster candidate c, we prompt the LLM to characterize it with a label. The label includes two components: an aspect and a brief cluster description. The aspect refers to a broad concept apparent in the top-5 talking points nearest the cluster centroid, and the description offers a short summary of these points.

The clustering output is based on conventional distance metrics, and it is not entirely perfect. Therefore, we perform a subsequent step to refine the set of cluster labels by merging redundant clusters and eliminating those that are inconsistent. To address redundancy, we greedily compare pairs of cluster labels and merge those that share the same aspect. The refined label set produced by removing redundancy represents the space of prominent points. Detailed procedures and hyperparameters for merging/removing are in Appendix E. The templates for the prompts used to characterize the candidate clusters and eliminate redundancy are shown in Tables 18 and 19, respectively.

Next, each talking point $t_s \in \mathcal{T}$ is assigned a cluster label by evaluating the cosine similarity between the embeddings of the cluster label and point t_s. This process results in a clustering $\{\mathcal{C}_j\}_{j=1}^{k}$ of talking points with associated cluster labels $\{\mathcal{L}_j\}_{j=1}^{k}$, which are termed the *prominent talking points*. More details are found in Appendix E.3.

Algorithm 1 Identify prominent talking points

Input: Talking points $\mathcal{T} = \{t_s\}_{s=1}^{p}$
Initialize: embeddings $\mathcal{Z} = \{z_s = f(t_s)\}_{s=1}^{p}\}$, n $\leftarrow$ no. of news articles, $\mathcal{C} \leftarrow \{\}$;
while $|\mathcal{Z}| > 0.1*$ n **do**
  clusters $\leftarrow$ *Clustering*($\mathcal{Z}$);
  labelSet $\leftarrow$ [];
  for c in clusters **do**
    Compute centroid μ_c by averaging;
    $Z' \leftarrow$ *getTopKPoints*(c, μ_c);
    cLabel $\leftarrow$ *getClusterLabel*(Z');
    Append cLabel to labelSet;
  end for
  updatedLabels $\leftarrow$ *updateLabelSet*(labelSet);
  $\mathcal{S} \leftarrow$ *TalkingPtMembership*($\mathcal{T}$, updatedLabels)
  $T' \leftarrow$ *getClusteredDocs*($\mathcal{S}$)
  $\mathcal{T} \leftarrow \mathcal{T} \setminus T'$;
  $\mathcal{Z} \leftarrow \mathcal{Z} \setminus$ {embeddings of T'};
  Append clusters in $\mathcal{S}$ to $\mathcal{C}$
end while
Output: k clusters $\mathcal{C} = \{\mathcal{C}_j\}_{j=1}^{k}$ with cluster labels $\{\mathcal{L}_j\}_{j=1}^{k}$

3.2.2 Generate Partisan Perspective

The clustering process produced a set of *prominent talking points* that are topically relevant to the event $\mathcal{E}$. However, these points lack ideologically specific insights crucial for defining partisan views. Consequently, in this step, we assign an ideology label to each talking point within the cluster and utilize the accompanying metadata to generate viewpoints specific to the *left* and *right* political ideologies. These viewpoints reflect how each political ideology discusses $\mathcal{E}$. Specifically,

we assign an ideology label {*left, right, center*} to each talking point within each cluster $\mathcal{C}_j$. Every news article is labeled with an ideology based on its media source, and each talking point, derived from these articles, inherits the corresponding label.

Next, we describe the method for creating the partisan viewpoint of the cluster by illustrating how the left-leaning summary is generated (the process for the right-leaning summary is analogous). Our aim is to produce a summary that distinctly represents leftist viewpoints, distinguishing it from the rightist perspective. To achieve this, after labeling the talking points within each cluster, we instruct the LLM to create the left-leaning summary in a contrastive style. We provide the LLM with characteristic talking points from both left and right ideologies so that it can highlight the contrasts and identify what characterizes the left viewpoint. Specifically, for a cluster $\mathcal{C}_j$, the prompt includes top-K talking points with a left bias alongside with top-M[2] talking points with a right bias for comparison. These representative points for each ideology are the nearest points to the cluster label selected based on the cosine similarity between the respective embeddings.

We note that the context provided by the top-K left-leaning points alone is limited and fails to capture the potential ideological bias exhibited by the entire article, which is necessary to generate a partisan summary. Ideally, the news articles linked to the top-K points should be used to contextualize them. However, due to the limitations in context length, we choose to use the article summaries.

We include the article summaries corresponding to each of the top-K left-leaning points in the prompt. To ensure these summaries reflect ideological biases and relevant information in the talking points, we instruct the LLM to generate the article summaries by conditioning the ideological label of the article and the aspect linked to the cluster $\mathcal{C}_j$.

Further, we want to ensure the generated left-leaning summary captures the relationship between the associated entities in the top-K left-leaning points. Therefore, we include metadata details such as actors, targets, sentiment on the targets, and the relevant media frame in the prompt. The template for generating the partisan summary is in Table 15.

[2] $M \leq K$; M & K are hyperparameters.

Issue	No. of Articles	No. of Events
Climate Change	579	8
Capitol Insurrection	1,609	4
Immigration	1,137	4
Coronavirus	2,816	8
Total Count	6,141	24

Table 1: The dataset we use for testing our proposed framework. It is sampled from Nakshatri et al. (2023).

4 Dataset

To illustrate the effectiveness of our proposed framework, we use the `keyevents` dataset (Nakshatri et al., 2023). This dataset is constructed by segmenting the archive of news articles from NELA-2021 (Horne et al., 2022) into a set of temporally motivated news events. To construct these events, Nakshatri et al. (2023) dynamically analyzed the temporal trend of news articles published for a given issue, and identified the temporal landmarks that could signify the presence of an important news event. Then, the news articles published in and around the temporal landmarks were clustered to identify all the documents relevant to the news event. In this manner, Nakshatri et al. (2023) proposed a dataset comprising of $40k$ news articles with 611 key news events from 11 issues.

As our goal is to analyze political discourse and characterize partisan perspectives at an event-level granularity, this dataset can be directly applicable to test the efficacy of our narrative framework. Thus, we manually sample a set of six issues and a set of events which have the highest number of news articles per event from this dataset. Table 1 shows the detailed statistics of our final dataset.

5 Evaluation

The effectiveness of our framework is verified by mainly using two kinds of automated evaluations (described below) in addition to human evaluation. Additional evaluation of intermediate framework steps is detailed in the Appendix A.3.

Ideology Classification. The main objective is to assess how well our framework can generate ideological discourse at event-level granularity. We test this through *ideology prediction* of unseen news articles related to the events in focus. If the generated ideological views effectively capture the broader discourse of the entire event $\mathcal{E}$, then using only the left and right perspectives, the LLM should correctly assign the ideological label to an unseen article from $\mathcal{E}$. This test is performed in various contexts, both with fine-tuning and without fine-tuning an LLM.

Partisan Classification. Ideology classification assesses the generated perspectives at the event level, whereas, through this task, we evaluate the "goodness" of these generated left- and right-perspectives at the PTP cluster level, as derived from Section 3.2.2. Specifically, within a cluster, we expect left viewpoints to indicate how *left* political ideology addresses the issue in relation to its cluster label (prominent point) and vice versa. Consequently, in an ideal scenario, left viewpoints should be supported by news articles from left-leaning sources and not reflected in articles from right-leaning sources (and similarly for the right). This task evaluates the degree to which the generated viewpoints exhibit this pattern.

6 Experiments & Results

In this section, we discuss the experimental details and findings linked to both automated and human evaluations. Initially, to generate ideology-specific viewpoints, we use ChatGPT [3] as the LLM to analyze each article, following the approach detailed in our proposed framework (Section 3). This process yields structured representations for articles from every event in our dataset, and we release these, along with the original dataset, to the community.

6.1 Ideology Classification Task

Task Formulation. Given an *unseen* news article that is not part of the initial clustering process but related to the event, the task is to predict the ideological label of the article - *{left, right}*. The central idea involves employing the generated partisan viewpoints for classification. We start by constructing the ideological discourse for the entire event by merging all the partisan viewpoints from each cluster (derived using our method), according to their ideology. We then use cosine similarity to identify the 3 most similar left- and right-viewpoints related to the input article, which we label as the *Event Partisan View*. With the input article and this *Event Partisan View*, we prompt the LLM to assign an ideology label to the article. In this process, we note that we mask the terms *{left, right}* and use *{summary1, summary2}* instead, to avoid any inherent bias from the LLM. The classification relies solely on the LLM's ability to align the article with the closest summary. For this task, we curated 481 unseen event-related articles (see Appendix D.3 for details).

[3] gpt-3.5-turbo-0125 (OpenAI, 2022)

Comparison Methods. We benchmark task performance with two LLMs, ChatGPT and LLAMA3[4] (Dubey et al., 2024), using these methods: **1. Direct Prompting**, where the LLM is directly prompted for the article's ideology label. **2. Event Partisan View**, and **3. Event View + Metadata**, which adds metadata to *Event Partisan View*. Metadata is aggregated from the top 50% of articles closest to the left- and right-viewpoints in each cluster (obtained from Section 3.2.2). Details on metadata construction can is in Appendix D.2.

Results. Table 2 presents ideology classification task results. *Event Partisan View* outperforms the direct prompting baseline, improving F1-scores by $+4.5$ for ChatGPT and $+1$ for Llama. This suggests that our framework's perspectives effectively capture event-level partisan signals that are essential for predicting the ideology of an unseen article. In addition, we observe that the inclusion of metadata can help improve classification task performance.

Narrative-LLAMA. To further show the effectiveness of generated partisan viewpoints from our framework, we use these viewpoints to fine-tune a LLAMA3-8B-instruct model. This is designed to directly generate high-level ideological viewpoints for the input article by bypassing the intermediate framework steps, utilizing PEFT [5] and LORA (Hu et al., 2021). We compiled 1100 examples pairing articles with their partisan perspectives derived from our framework, and trained the model using DPO (Rafailov et al., 2024) to align generated viewpoints with article's inherent ideological bias (model training details can be found in Appendix D.4). In this case, classification is assessed via: **Prompting with Generated Perspective**, obtaining perspectives with Narrative-LLAMA and prompting for article ideology.

Discussion. From Table 3, it is evident that the fine-tuned Narrative-LLAMA outperforms the LLAMA3-8b-instruct model by increasing the F1 score by $+3.6$ points in the ideology classification task. This improvement reflects the effectiveness of our framework's training data in capturing ideology-specific nuances within event-discourse. Furthermore, we assess the quality of the generated perspectives through an automated process using the Llama3-70B-instruct model. This validates if the generated perspectives actually appear in the

[4] meta-llama/Meta-Llama-3-8B-Instruct
[5] https://github.com/huggingface/peft

LLM	Method	Prec.	Recall	F1
ChatGPT (0-shot)	Direct Prompting	81.38	74.66	73.52
	Event Partisan View	77.12	76.64	76.61
	+Metadata	81.20	79.78	**79.69**
ChatGPT (2-shot)	Direct Prompting	80.47	78.23	78.09
	Event Partisan View	83.02	82.61	**82.65**
	+Metadata	83.79	82.50	82.52
LLAMA (2-shot)	Direct Prompting	79.54	81.0	79.77
	Event Partisan View	80.45	81.71	80.78
	+Metadata	81.62	82.62	**81.93**

Table 2: Averaged results for ideology classification across 4 issues. Generated *Event Partisan View* effectively captures event-level ideological narratives.

LLM	Method	Prec.	Recall	F1
LLAMA	Prompting w Generated Perspective	87.8	84.9	84.9
Narrative-LLAMA	Prompting w Generated Perspective	89.5	88.4	**88.5**

Table 3: Narrative-LLAMA (fine-tuned model) improves over Llama3 (+3.6 F1-score) in ideology classification.

original news articles, as detailed in appendix A.2. Notably, hallucination rates were found to be below 6%, and we intend to release the model parameters to the community. Finally, manual inspection of the perspectives generated by Narrative-Llama reveal that they tend to capture broader discourse, rather than focusing on article-specific details. Figure 5 provides a relevant example.

6.2 Partisan Classification

Task Formulation. The ideology classification task evaluates generated discourse at the event-level using unseen news articles, whereas this task assesses the partisan behavior of the generated left- and right-viewpoints at the cluster level (obtained from Section 3.2.2). Each cluster includes left- and right-leaning talking points. We consider the related news article for each such talking point in the cluster. In the context of a left-leaning article along with its left- and right-leaning perspectives in the cluster, the goal is to assign the news article to the summary it aligns with more closely. Successfully achieving this implies that the generated perspectives reflect particular ideology-specific elements, thereby exhibiting partisan behavior at the cluster level. Following our earlier discussion in task 6.1, in this task as well, we replace the terms *{left, right}* with *{summary1, summary2}*. This substitution is intended to prevent any bias from the LLM. Prompt template shown in Table 17 in the appendix.

Experimental Setup. We benchmark the task performance using ChatGPT, using the following methods: **1. Partisan View**; represents the perspectives generated by our framework 3.2.2. **2. Partisan View + Metadata**; adds metadata aggregated from the top 50% of articles closest to left- and right-viewpoints in the cluster (more details can be found in Appendix D.2). **3. Topically Relevant Points (TPs)**; as a baseline measure, in each cluster, we consider the top-3 talking points nearest to the prominent talking point (or cluster label), as identified in Section 3.2.1, prior to generating left- and right-perspectives. These top-3 TPs are obtained for each ideology label, and we use them as left- and right-perspectives.

Discussion. Table 4 shows the results of partisan classification, with issue-specific results available in Appendix A.1. The talking points from the clustering process (3.2.1) lack the ability to adequately capture ideological nuances in political discourse, as they underperform *Partisan View* by 11 F1-score points. In contrast, *Partisan View* consistently enhances the F1-score for all issues, confirming its efficacy in demonstrating partisan behavior at the cluster-level essential for understanding ideological discourse. Additionally, incorporating extracted metadata adds approximately +1 point to the F1-score, further enhancing performance.

Approach	Prec.	Recall	F1
TPs	73.44	73.33	73.37
Partisan View	85.03	84.61	84.76
Partisan View + Metadata	**85.93**	**86.14**	**85.98**

Table 4: Averaged results for *partisan classification task* across all issues shows the efficacy of partisan perspectives in capturing ideology-specific information.

6.3 Human Evaluation

To assess our framework's ability to generate left- and right-viewpoints, we conduct human evaluations on a small data sample (comprised of 3 events listed in Table 11 in the appendix) and present our results. The focus is on testing ***summary coherence*** and ***mapping quality***. The ideological viewpoints in 3.2 are primarily based on the top-K talking points within each cluster. For *summary coherence*, we examine if the generated perspectives accurately reflect the key ideas from the top-K talking points and coherently present them. For *Mapping Quality*, we verify whether the left- and right-perspectives are properly expressed in corresponding articles (see Table 5). Evaluating mapping quality on a large scale is challenging due to the need to read lengthy article excerpts. To ad-

dress this, we also consider ***MQ_LLM***, leveraging GPT-4o to extract relevant evidence from news articles for questions related to the topic, positively and negatively viewed entities, etc. This evidence is then validated by humans rather than reviewing entire articles. We perform this analysis for a Climate Change event (see Table 6). Details about the experimental setup can be found in Appendix A.4.

Discussion. Table 5 reveals that generated partisan viewpoints, for both political ideologies, have high coherence scores, indicating that they align well with talking points and the article summaries from which they were created. High mapping quality scores suggest that these summaries are accurately reflected in news articles. Upon manual review, we found that the generated partisan perspectives can be incorrect, particularly when the LLM generates poor summaries of news articles that are used for creating these perspectives, as shown in Table 21. In addition, the LLM occasionally overlooks the information cited in the news articles, leading to incorrect summaries. Finally, Table 6 shows a high MQ_LLM agreement score between GPT-4o and humans, with least score for when entities are viewed negatively. In this particular case, manual inspection shows that GPT-4o sometimes misses relevant evidence from the articles, exemplified in Table 12 in the Appendix.

Issue	L-Coherence	R-Coherence	L-MQ	R-MQ
Climate Change	85.71	100	75.00	76.92
Coronavirus	100	90.90	90.90	70.00
Immigration	93.33	100	84.62	94.44

Table 5: Results from 84 annotations indicate that partisan perspectives are indeed expressed in the original news articles. **L** refers to left political ideology. **R** refers to right political ideology.

7 Broader Impact

Analyzing Event-Level Narratives. With our framework, political scientists can understand how each political party discusses an event without analyzing every relevant article. We introduce ***event-discourse snapshots***, a visual approach to examine event-level narratives. Figure 4 illustrates a *snapshot* for an event concerning Coronavirus - *Biden's COVID-19 Relief Bill*. Each generated perspective (circle) on the X-axis reflects its bias, with balanced view placing it in the center. Circle size indicates its frequency. Highly repeating perspectives with high positive or negative x-axis values is evidence for "different realities", i.e., focusing on very different aspects of the topic. This visualization enables event-level discourse analysis by sorting perspectives into types like *agreement, disagreement, one-sided, etc.* Table 13 provides an example for the same event. Details of visual creation are in Appendix A.5.

Question Type	Agreement Score
Topic	97.82
Entities viewed negatively	80.43
Entities viewed positively	92.39
Angle of discussion	98.91

Table 6: Shows the GPT4o-Human agreement score for 92 article-perspective pairs (for *Climate Change* related event).

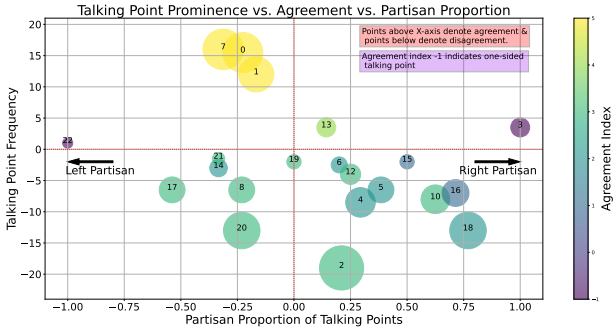


Figure 4: Shows talking point prominence vs. points agreed/disagreed vs. partisan proportion (for an event related to Coronavirus).

Temporal Analysis. Our framework can be easily applied to examine temporal trends of talking points across events related to an issue, emphasizing parts of events that interest political parties (refer to Appendix B). Additionally, the variation in metadata over time helps in understanding partisan perspectives on an issue. Figure 6 illustrates this temporal trend for the issue *Covid*.

8 Conclusion

In this paper, we introduced an LLM-based framework for analyzing ideological discourse in news events. We demonstrated the effectiveness of our approach through both automated tasks— ideology and partisan classification, and human validation. Additionally, we showcased the basic applications of our framework in event-discourse analysis, with plans for further investigation into its broader applicability in future work. Notably, our human evaluation revealed that LLMs struggle to accurately interpret articles citing multiple sources, which introduces noise into discourse analysis. We aim to explore this in future research.

9 Limitations and Ethics

We identify the following limitations from our study.

Closed-source LLM. As a first, the perspectives generated from our framework are based on an LLM model, ChatGPT, which is closed source, and the details of its construction are unclear (Spirling, 2023). We chose this model because it delivered higher quality output than other options available at the time.

Talking point extraction. Our framework allows the LLM to decide the key talking points from the news article, although it is possible that it could overlook a prominent talking point. While this is a potential limitation, we believe that if a talking point is really prominent, then it will repeat in many articles, to shape the narrative. Thus, there is a high chance that the LLM would identify that talking point in other articles, even if the model failed to recognize the prominent point in the given article.

Labeling the talking point. Further, we assume that all the talking points from a left-leaning news source are actually *left-biased*, and vice-versa. However, in reality, it need not be the case (Kim et al., 2022). Our approach performs fairly well, even with this assumption primarily because we are only interested in identifying salient talking points from each ideology, and less frequent talking points are rejected.

Cautious use of our framework. We acknowledge that our framework produces partisan viewpoints through a language model, which can occasionally generate inaccurate or misleading representations, often referred to as LLM-hallucinations. These inaccuracies may not be easily identified through automated evaluation methods. Consequently, we strongly advise users to seek expert validation before using our framework.

While our system offers numerous practical applications, we urge users to approach its usage with caution. Although it can effectively identify ideologies, there is a potential for misuse, such as targeting individuals based on their expressed beliefs or affiliations. For these reasons, as well as many others, it is crucial for users to thoroughly evaluate both the advantages and potential risks associated with deploying our framework.

Ethics. To the best of our knowledge, we did not violate any widely held ethical precepts when producing this paper. All results are from a machine learning model and should be interpreted as such. We present all implementation and dataset details for reproducibility of our study (some parts are in the appendix). The datasets and LLMs used in this paper are publicly available and allowed for scientific research.

References

Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624, Online. Association for Computational Linguistics.

Amber E. Boydstun, Dallas Card, Justin H. Gross, Paul Resnick, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015a. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015b. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Dennis Chong and James N Druckman. 2007. Framing theory. *Annu. Rev. Polit. Sci.*, 10:103–126.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Elizabeth Dubois and Grant Blank. 2018. The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, Communication & Society*, 21(5):729–745.

Robert M. Entman. 1993. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, 43(4):51–58.

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.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. In *2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lea Frermann, Jiatong Li, Shima Khanehzar, and Gosia Mikolajczak. 2023. Conflicts, villains, resolutions: Towards models of narrative media framing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*, pages 913–922. International World Wide Web Conferences Steering Committee.

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

Matthew Gentzkow and Jesse M Shapiro. 2011. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839.

Erving Goffman. 1974. *Frame analysis: An essay on the organization of experience.* Harvard University Press.

Benjamin Horne, Mauricio Gruppi, and Sibel Adali. 2022. NELA-GT-2021.

Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2023. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13188–13214.

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. *arXiv preprint arXiv:2106.09685*.

Kathleen Hall Jamieson, Bruce W Hardy, and Daniel Romer. 2007. The effectiveness of the press in serving the needs of american democracy.

Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005, Vancouver, Canada. Association for Computational Linguistics.

Shima Khanehzar, Trevor Cohn, Gosia Mikolajczak, Andrew Turpin, and Lea Frermann. 2021. Framing unpacked: A semi-supervised interpretable multi-view model of media frames. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2154–2166, Online. Association for Computational Linguistics.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Eunji Kim, Yphtach Lelkes, and Joshua McCrain. 2022. Measuring dynamic media bias. *Proceedings of the National Academy of Sciences*, 119(32):e2202197119.

Donald R Kinder. 1998. Communication and opinion. *Annual review of political science*, 1(1):167–197.

Andrea Lawlor and Erin Tolley. 2017. Deciding who's legitimate: News media framing of immigrants and refugees. *International Journal of Communication*, 11(0).

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. Neus: Neutral multi-news summarization for mitigating framing bias.

Chang Li and Dan Goldwasser. 2019. Encoding social information with graph convolutional networks forPolitical perspective detection in news media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604, Florence, Italy. Association for Computational Linguistics.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Siyi Liu, Hongming Zhang, Hongwei Wang, Kaiqiang Song, Dan Roth, and Dong Yu. 2023. Open-domain event graph induction for mitigating framing bias.

Y Liu, X Zhang, D Wegsman, N Beauchamp, and L Wang. 2022. Politics: Pretraining with same-story article comparison for ideology prediction and stance

detection. *Findings of the Association for Computational Linguistics: NAACL 2022*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315.

Maxwell E McCombs and Donald L Shaw. 1972. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263.

Davoud Moulavi, Pablo A Jaskowiak, Ricardo JGB Campello, Arthur Zimek, and Jörg Sander. 2014. Density-based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 839–847. SIAM.

Nishanth Nakshatri, Siyi Liu, Sihao Chen, Dan Roth, Dan Goldwasser, and Daniel Hopkins. 2023. Using llm for improving key event discovery: Temporal-guided news stream clustering with event summaries. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4162–4173.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. 2021. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*.

OpenAI. 2022. GPT-3.5 (ChatGPT). Computer software.

Kunwoo Park, Zhufeng Pan, and Jungseock Joo. 2021. Who blames or endorses whom? entity-to-entity directed sentiment extraction in news text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4091–4102.

Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo chambers on facebook. *Available at SSRN 2795110*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321.

Shamik Roy, María Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Dietram A Scheufele and David Tewksbury. 2007. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication*, 57(1):9–20.

Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. 2021. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3):102505.

Arthur Spirling. 2023. Why open-source generative ai models are an ethical way forward for science. *Nature*, 616:413.

Teun A Van Dijk. 1996. Ideological discourse analysis. *MOARA–Revista Eletrônica do Programa de Pós-Graduação em Letras ISSN: 0104-0944*, (06):13–45.

Teun A Van Dijk et al. 1997. What is political discourse analysis. *Belgian journal of linguistics*, 11(1):11–52.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, pages 1–55.

A Extended Results

A.1 Partisan Classification

Table 7 shows the partisan classification results for each issue and the total aggregated performance.

A.2 Ideology Classification Task

Narrative-LLAMA Quality Test. We assess whether Narrative-LLAMA's generated perspectives are reflected in the input news articles. To achieve this, we use an economical automated evaluation approach with a *LLAMA3-70B-instruct* model. Specifically, for all 481 unseen input articles, we first generate high-level ideological viewpoints using Narrative-LLAMA. We then prompt LLAMA3-70B-instruct to check if these viewpoints are represented in the corresponding news article (for quality). We use two metrics for success: **hard** and **soft**. For each article, the **hard** metric scores 1 if *all* generated viewpoints are reflected in the article (indicating no hallucination). The **soft** metric calculates the overall proportion of viewpoints conveyed in the input articles. Table 8 presents these findings per issue, revealing that on an average, in 94% of cases, *all* generated perspectives are included in the input news articles, as evaluated by the 70B-instruct model.

A.3 Evaluate the prominent talking points

We evaluate our framework's ability to effectively cluster the talking points using two metrics - *coverage*, and *topic diversity*. In Section 3.2.1, we had built a set of *prominent talking points* by clustering the set of all points $\mathcal{T}$ associated with the event $\mathcal{E}$. These *prominent talking points*, which capture topically relevant information, characterize the space of repeating talking points for that event.

Coverage. We define coverage to be the proportion of talking points that were taken into account in the clustering process. A 100% coverage would imply that we are accounting for all the information in the input talking points. Therefore, if the *prominent talking points* set are a good representation for all the clusters, then we expect that each talking point in $\mathcal{T}$ should be able to map back to one of the *prominent talking points*. Table 9 shows the average coverage for each issue. We observe that identified *prominent talking points* cover at least 80% of the talking point set $\mathcal{T}$, indicating that they are a good representation of the set $\mathcal{T}$.

Topic Diversity. We also validate if the *prominent talking points* capture diverse topics.

Task formulation. To this end, we formulate the following ***topic classification task*** : Given a talking point and a set of K' cluster labels, assign the talking point to the most topically relevant cluster label k^*, where $k^* \in K'$. Note that the talking point is associated with only one of the K' labels, and the rest of the labels are randomly sampled negative examples (other clusters that don't have the talking point). The negative examples help assess the degree of cluster separation. Precisely, k^* helps assess how well the talking point assignments to map to their respective clusters, whereas the remaining negative labels, $K' \setminus \{k^*\}$, help measure the degree of separation between the clusters.

Experimental Setup. We first split the data in each cluster into 4 quartiles, where the 1st quartile refers to the top 25% closest talking points (TP) to the corresponding cluster label in the embedding space, the 2nd the top 50%, etc. We randomly sample half the TPs from each quartile for this experiment, with 3 neg. labels for each TP ($|K'| = 4$). We prompt ChatGPT to assign the TP to its most topically relevant label (prompt: Table 16).

Table 10 shows the performance of the topic classification task. We see all quartiles perform well, and documents closer to the cluster label (lower quartile) show strongest topical relevance to the cluster label. This shows that our cluster labels do clearly capture the diverse topics of our talking points, and each cluster captures a unique aspect, when compared to other clusters.

Discussion. Topic classification results indicate that topics associated with *prominent talking points* are diverse ($\approx 80\%$ accuracy) when compared with one another, while the coverage indicate that the prominent points span at least 80% of the set $\mathcal{T}$. On combining both these dimensions, we observe that our approach forms reasonable set of *prominent talking points*.

A.4 Human Evaluation

We perform human evaluation on a set of three events for three different issues. In this section, we describe the annotation procedure for each task. Note that annotations were performed for a total of 84 talking points across three issues for the metrics - *Summary Coherence*, and *Mapping Quality*.

Issue	Approach	Avg. Precision	Avg. Recall	Avg. F1-score
Climate Change	Topically Relevant Points	84.11	84.23	84.17
	Partisan View	91.73	89.46	90.29
	Partisan View + Metadata	92.43	90.86	**91.49**
Capitol Insurrection	Topically Relevant Points	69.50	71.62	69.18
	Partisan View	79.33	80.93	**79.93**
	Partisan View + Metadata	81.04	78.08	79.12
Immigration	Topically Relevant Points	69.14	74.64	69.92
	Partisan View	85.38	86.36	85.85
	Partisan View + Metadata	88.27	86.17	**87.15**
Coronavirus	Topically Relevant Points	73.11	72.60	72.77
	Partisan View	83.34	81.76	82.21
	Partisan View + Metadata	83.78	84.20	**83.92**
Overall Performance	Topically Relevant Points	73.44	73.33	73.37
	Partisan View	85.03	84.61	84.76
	Partisan View + Metadata	85.93	86.14	**85.98**

Table 7: Averaged results for *partisan classification task* shows the efficacy of partisan perspectives in capturing ideology-specific information at the cluster-level granularity.

LLAMA:
1. Joe Biden's Executive Orders: The article reports that President-elect Joe Biden plans to sign a raft of Executive Orders on his first day in office, covering issues such as COVID-19, climate change, racial equity, and economic recovery. These actions aim to address the four crises identified by Biden's transition team.

2. Biden's Priorities: The article highlights Biden's focus on addressing the COVID-19 crisis, economic recovery, climate change, and racial equity. His administration plans to take decisive action to deliver relief to millions of Americans struggling with these crises.

ChatGPT:
1. Biden's Executive Orders: President-elect Joe Biden plans to sign several Executive Orders during his first days in office, addressing key issues such as COVID-19, climate change, student loan relief, and reversing Trump-era policies.

2. Focus on Four Major Crises: Biden's administration will prioritize tackling four major crises—COVID-19, economic recovery, climate change, and racial equity—through immediate executive actions.

Narrative-LLAMA:
1. Biden's Executive Order Overreach: Biden's Use of Executive Orders to Implement His Agenda without Congressional Input\nThe article criticizes President-elect Biden's plan to implement his agenda through Executive Orders, bypassing Congress and potentially reversing Trump's policies, highlighting concerns about Biden's willingness to circumvent legislative process and the potential impact on his ability to implement his agenda effectively.

Figure 5: Shows an example of the generated perspectives for a *right*-leaning news article related to *Biden's executive orders*. The generated perspectives from Narrative-LLAMA capture broader ideological discourse as opposed to its counterparts.

Issue Name	% Success Rate	
	Soft	Hard
Climate Change	97.70	93.53
Capitol Insurrection	95.97	92.15
Immigration	98.28	95.73
Coronavirus	98.39	95.06

Table 8: LLAMA-3-70B-instruct evaluates issue-specific mapping quality, finding that on average, 94% of ***all*** the generated viewpoints are found in the input news articles.

Issue	Avg. Coverage per event	Avg. # clusters
Climate Change	83.17	10
Capitol Insurrection	86.70	24
Immigration	90.55	21
Coronavirus	78.18	16

Table 9: Averaged results for coverage.

Summary Coherence. We explain the procedure for the *left* political ideology, and the same process is repeated for the *right* ideology as well. First, we explain the task to the annotators with an example. The annotators are provided with a left-perspective along with three to five left talking points and their respective news article summaries. We ask the annotators to validate if the left-perspective can be derived from the news article summaries or the talk-

Issue	Q1	Q2	Q3	Q4
Climate Change	91.19	87.47	83.66	80.00
Capitol Insurrection	91.78	89.34	84.56	80.27
Immigration	91.96	88.69	85.01	80.34
Coronavirus	94.07	89.11	84.10	79.94
Avg. Accuracy	**92.74**	**88.90**	**84.37**	**80.12**

Table 10: Averaged results for each quartile for the *topic classification task* indicates that our prominent talking points capture diverse information.

ing points. If it can be derived, then the response is 1, otherwise it is 0. In the cases where annotators are not sure, the response is -1.

Mapping Quality. We explain the procedure for the *left* political ideology, and the same process is repeated for the *right* ideology as well. In this case, we provide the annotators with a left-perspective and a corresponding news article that is most relevant to the left-perspective (measured based on cosine similarity distance in the embedding space). We segment the news article into sentences of 7, and we only provide the most relevant 7 sentences from the news article to the annotators. First, we let the annotators know that there are at most three points in the left-perspective, then ask them to compare the left-view with the left news article content to validate if at least one of the points in the summary is expressed in the article. If it is, then the response is 1, otherwise it is 0. In cases where annotators are not sure, the response is -1.

Mapping Quality with GPT-4o (MQ_LLM). For a climate change-related event, we examined clusters from our framework (see Section 3.2). From each cluster, we randomly selected up to 5 articles for each ideological viewpoint (*{left, right}*), resulting in 92 article-partisan perspective pairs (comprising 48 *left* pairs and 44 *right* pairs). For each pair, GPT-4o was tasked with quoting relevant sentences from the articles to answer specific questions.

1. Is the summary discussing the same topic as the news article?
2. In the summary and the news article, are there any entities in common that are viewed negatively from the same perspective?
3. In the summary and the news article, are there any entities in common that are viewed positively from the same perspective?
4. Does the news article cover the views presented in the summary from the same angle?

For each article-partisan summary pair, we present the partisan view and evidence from the article (provided by GPT-4o) to a human. Human is expected to validate each answer by verifying if the retrieved evidence aligns with the summary. Note that we proposed this experiment as an alternative evaluation scheme for mapping quality, and we limit the scale to one event considering the API cost associated with GPT-4o model.

Note that our annotators were graduate STEM-students who were not the authors of the paper and were under the age of 30.

A.5 Constructing Visualization for Partisan Narrative

To visualize the partisan narrative for an event, we would need to obtain agreement/disagreement between the talking point perspectives - *left/right*. To obtain this, we define a scale, where we prompt a better-performing model, GPT-4o, to assign a binary label - $0/1$, for each of the following questions.

1. Do both summaries have at least one common aspect of discussion?
2. Are the summaries discussing about similar entities?
3. Are the entities in common viewed in the same manner? For example, is the entity viewed positively or negatively in both the summaries?
4. Do both the summaries talk about the event from the same perspective?
5. If the summaries are viewing the event from different angles, do the summaries have atleast some agreement with each other?

We obtained 2 talking points with a cumulative score of 1; 5 points with a score of 2; 8 points with a score of 3; 4 points with a score of 1; and 3 talking points with a cumulative score of 5. We note that higher scores indicate that talking points are closer to being in agreement with each other, whereas lower scores imply that talking points are mostly disagreeing with each other. For a score of 3, we manually inspected the outputs from the model and deduced that the two summaries shared a common aspect, discussed similar entities and had some agreements with each other. However, the entities were not viewed in the same manner due to which we assigned these talking points to disagree with each other.

Issue	News Event
Climate Change	**Event Title:** Biden Announces Ambitious Greenhouse Gas Emissions Cut **Event Description:** This is about President Joe Biden's announcement of an ambitious cut in greenhouse gas emissions as he looks to put the US back at the center of the global effort to address the climate crisis and curb carbon emissions.
Coronavirus	**Event Title:** Biden's COVID-19 Vaccination Mandate **Event Description:** This is about President Joe Biden's announcement of new COVID-19 vaccination requirements for federal government employees, healthcare workers, and companies with 100 or more employees, and his criticism of politicians who are undermining trust in COVID vaccines.
Immigration	**Event Title:** Biden's Refugee Cap Decision **Event Description:** This is about the criticism faced by President Biden for his decision to not raise the cap on refugees allowed to enter the US this year, which he had promised to do during his campaign.

Table 11: Events considered for human evaluation.

Talking Point Summary	Evidence From Article
Uncertainty in global cooperation and skepticism towards US leadership.Concerns persist over the uncertainty of international support, especially from major carbon emitters like China, India, and Russia, towards America's climate initiatives. Differing views on the urgency of climate action and skepticism towards US leadership may hinder effective global collaboration on climate change.,	Evidence: Both the summary and the news article mention skepticism towards US leadership and the challenges in global cooperation. The summary states, "Concerns persist over the uncertainty of international support, especially from major carbon emitters like China, India, and Russia, towards America's climate initiatives." The news article similarly notes, "Russian President Vladimir Putin and Chinese President Xi Jinping are two notable leaders who have both confirmed their attendance at the summit, underscoring the wide range of leaders attending," indicating the importance of their participation and potential skepticism.

Table 12: GPT-4o fails to correctly identify the evidence from the news article.

B Temporal Analysis - Case Study

Here, we provide a simple case study to show how the data obtained from our framework could be utilized to study the *left* and *right* ideology viewpoints for an entire issue. To do this, we consider 7 events related to the issue *Coronavirus* at various points in time, and at every point, analyze the most frequently repeating prominent talking point from each political ideology.

Figure 6(a) shows a dynamic evolution of prominent talking point of each political party for the issue - *Coronavirus*. We observe that frequently discussed prominent point of each political party is different from one another in 3 out of the 7 events under consideration. However, both political parties predominantly discuss the same prominent point in the remaining cases. Note that Figure 6(a) shows only the *aspect* associated with each prominent point for data visualization clarity.

In the cases where both political parties discuss the same prominent point, we can further investigate the manner in which they talk about the prominent point by observing its corresponding partisan summary. For instance, let us consider the prominent point with the aspect - *Evolving mask guidelines post-CDC update*, that is commonly discussed by both political parties. While both the parties criticize the ambiguity in CDC's mask guidance, the *left*-leaning articles emphasizes more on pointing out the discrepancies with state and local mandates, and how it is impacting businesses. However, *right*-leaning sources focus on delayed response by CDC in updating mask mandates for vaccinated individuals and raises concern about the leadership.

We can further analyze this prominent point discussed by both parties through its associated metadata. The entity viewed as a *target* by an ideology, its corresponding *actor*, and the associated *media frame* can help analyze the differences in the viewpoints across political parties. For the same prominent point with the aspect *Evolving mask guidelines post-CDC update*, we observed that *left*-leaning

TP Type	TP ID	Left View (only titles)	Right View (only titles)
Agreement	1	Rejection of splitting COVID relief bill into separate components	Resistance to breaking down relief package into separate bills
Disagreement	10	Emphasis on Transparency and Improved Vaccine Distribution	Questioning Biden's Vaccine Distribution Transparency
Agenda Setting	16	Biden's Travel Restrictions and Bans for Public Health	Criticism of Biden's Executive Order on Pandemic Language (for banning term - 'China Virus')
Partisan Battle	2	Biden administration's emphasis on equitable vaccine distribution and healthcare reform	Criticism of Biden administration's vaccine distribution decisions
Right Only	3	-	Economic Impact of $15 Minimum Wage

Table 13: Overview of the talking points (TPs) based on its potential type.

news sources viewed the entity *Centers for Disease Control and Prevention (CDC)* to have negatively impacted the target entity *Retailers*. Further investigation revealed that it was due to the criticism associated with changing mask guidelines, where *CDC* removed mask mandates for the vaccinated individuals, and left-leaning sources criticized *CDC* for creating ambiguity amongst the *retailers* regarding the mask guidelines. We note that *left*-leaning news sources commonly used *Policy* as the media frame of discussion in the context of this actor-target pair. In this way, the metadata associated with the prominent point of interest can further help distinguish left and right perspectives. To obtain an overall global view of variation in metadata for the entire issue, Figure 6(b) shows a dynamic analysis over the actor/target entities for each prominent point across the two political parties over time.

C Schema

D Experiments Related

D.1 Prompt Templates for experiment section

D.2 Metadata generation

The goal from this step is to identify the most frequent and discriminative pair of entities along with their sentiments that can help distinguish between the two political ideologies. Note that for an ideology, we use only the top-K documents and its associated metadata for generating the partisan perspective. To account for the metadata from the rest of the members in the cluster and obtain a global cluster-view, we aggregate this information from the top-50% of the members in the cluster. Specifically, we obtain the top-3 target entities that have positive sentiment, and top-3 target entities that have negative sentiment. In each case, we obtain the most common actor associated with the respective target. We also obtain the most common mediaframe associated with the corresponding actor-target context. This information can be plugged into the prompt in addition to the partisan viewpoints to help better distinguish between the political ideologies.

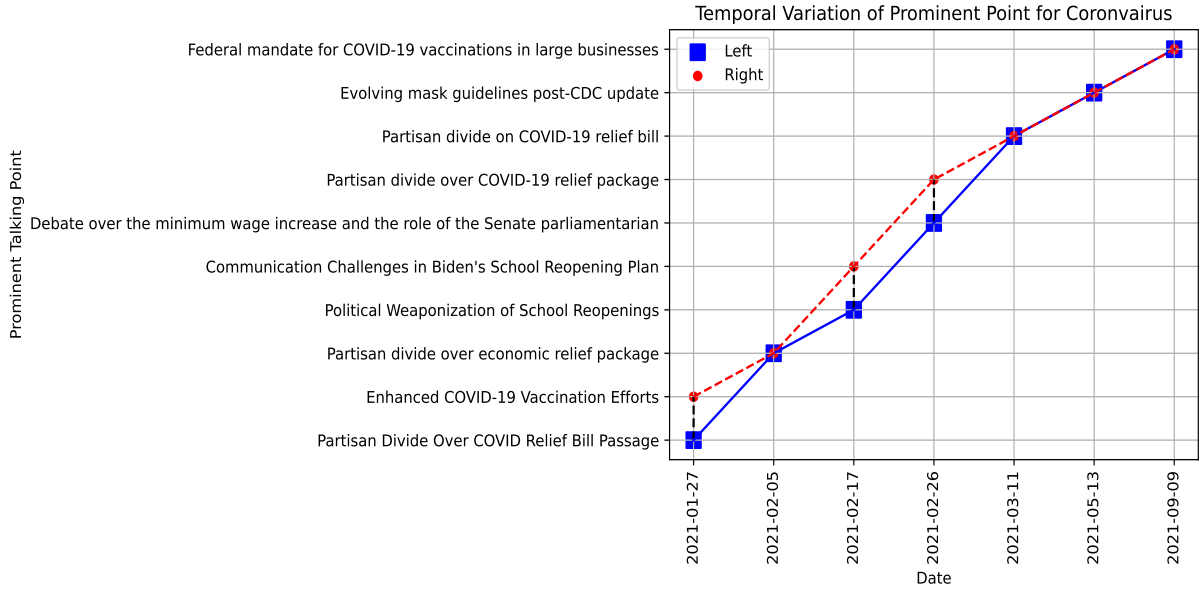
D.3 Dataset extraction

Here, we describe the process used for extracting the set of *unseen* news articles. We note that (Nakshatri et al., 2023) used NELA-2021 dataset for segmenting the news articles into a set of temporally motivated news events. In this process, (Nakshatri et al., 2023) used a temporal window of 3 in order to obtain coherent news events.

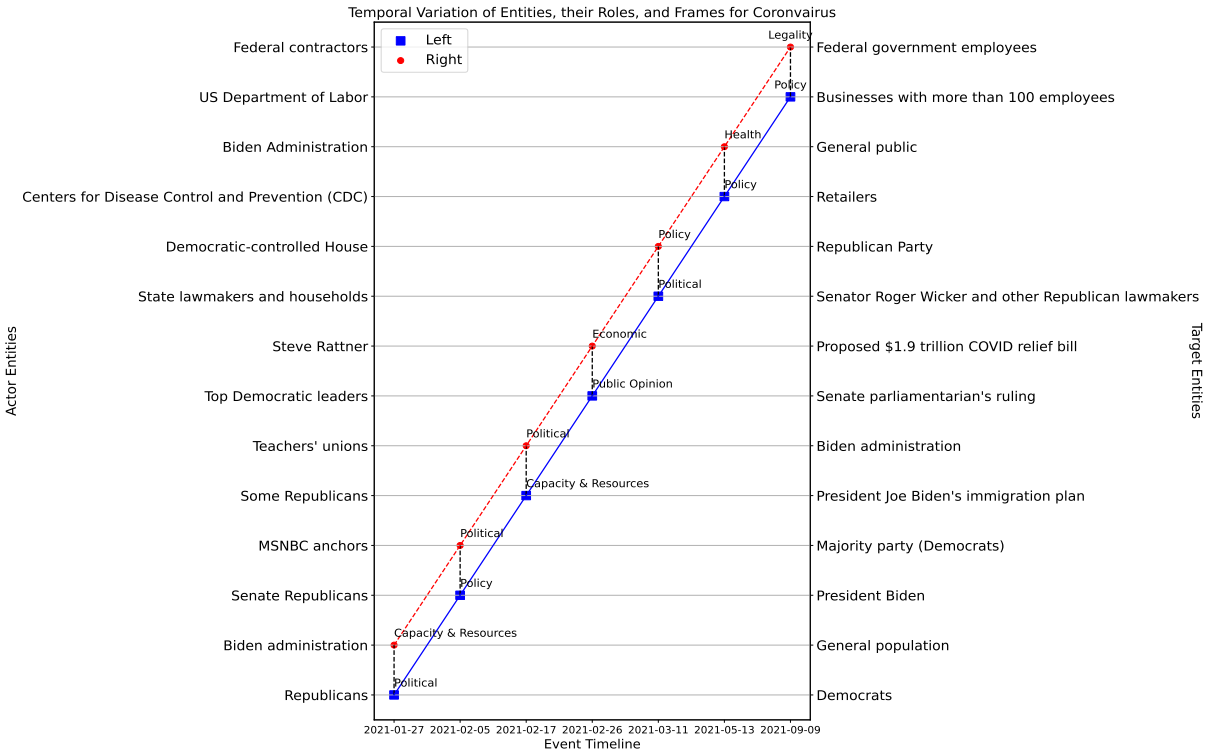
In order to obtain *unseen* news articles, yet relevant to the events under consideration, we extend this temporal window to 7 days, and retrieve all the news articles for that time period from NELA-2021 dataset. We filter out all the news articles that part of our clustering process. Then, we consider the all the unseen articles that are closest to the event centroid in the embedding space (threshold ≥ 0.86). Note that we obtain event centroid by averaging the embedding of all news articles relevant to the event. In this way, we extracted 481 relevant news articles for the events under consideration, of which 234 news articles are from right-leaning news sources, and the rest are from the left-leaning news sources.

D.4 Narrative-LLAMA Training Details

Our goal is to develop a model capable of capturing higher-level ideological views within news articles, rather than focusing on minor details that are not apparent in the broader event discourse. To eliminate these minor details and enable the model to focus on the event-level ideological discourse, we train an open-sourced LLAMA3-8B model with the partisan viewpoints generated by our proposed framework. Specifically, we fine-tune the LLAMA-8B-instruct model using news article as input, and its corresponding partisan viewpoint as output. In this process, we observe that the partisan perspectives provided in training are from our framework and generated by ChatGPT, a larger LLM with superior reasoning ability compared to LLAMA3-8B. Therefore, in a way, we are distilling knowledge from a larger model (ChatGPT) to



(a) Compares the temporal variation of most frequent prominent point for each political ideology, and across 7 events related to the issue - *Coronavirus*. X-axis indicates event-timeline, and Y-axis shows the most prominent talking point label. Frequently discussed prominent point across the two ideologies intersect in 4 out of 7 cases.



(b) Illustrates the temporal variation in the metadata formatted as *actor-negatively impacting-target*. We show frequently repeating target entity with a negative sentiment for each ideology, and across 7 events for the issue - *Coronavirus*. For each target entity, its corresponding actor, and the associated media frame is also shown.

Figure 6: Temporal analysis of prominent points along with its respective metadata for the issue - *Coronavirus*.

Schema	Schema Prompt Template (incremental)
Talking Point	You are tasked with discerning the key talking points from the following 'NEWS ARTICLE'. Your objective is to condense the contents of the 'NEWS ARTICLE' into a succinct list of up to four primary talking points, each accompanied by a brief description. \n 'NEWS ARTICLE': {article}
Entities	For each 'Talking Point', identify up to three pivotal entities associated with it, and assess whether the author of the 'NEWS ARTICLE' exhibits a bias either against or in favor of the mentioned entities. In instances where no discernible bias is evident, categorize the entity as neutral. The goal is to compile a list of entities along with their entity types, categorizing them as either against, in favor, or neutral, and accompany each categorization with a brief explanation.
Activity	For each 'Talking Point' and its associated 'Entities', identify the primary activities linked to it. For every identified 'Activity', pinpoint the entity assuming the role of 'Actor', driving the said 'Activity', and the entity acting as the 'Target', which is influenced by the 'Actor'. Assess whether the impact on the 'Target' is positive, negative, or neutral, providing a rationale for the impact. Focus only on pivotal 'Activities' closely related to the 'Talking Point'.
Media Frame	For each 'Talking Point' and its corresponding 'Activity', predict its media frame, and categorize it into one of 15 labels: Economic, Capacity & Resources, Morality, Fairness & Equality, Legality, Policy, Crime, Security, Health, Quality of Life, Cultural, Public Opinion, Political, External Regulation, or Other. With respect to the predicted 'Frame', provide a short explanation on how it is related to the main 'Activity'.

Table 14: Prompt template used to obtain the structured representation of the article along with the relevant metadata.

Generate Partisan Summary
You are provided with an aspect of discussion related to a news event, along with biased talking points from left and right political ideologies discussing the same aspect. Each talking point is associated with its respective news article summary, and metadata that includes actions, actors, targets, impacts, and framing. On comparing and analyzing the talking points from both ideologies, the objective is to refine and condense left-biased talking points into at most three unique points, such that the new points clearly capture the political bias towards left ideology. Redundant points and those not aligning well with left political ideology should be excluded. #### Input #### Aspect of discussion: {aspect} Left-biased talking points: {left-biased points} ## includes metadata for each point Left-biased news article summaries: {left-biased summaries} Right-biased talking points: {right-biased points}

Table 15: Prompt template used to obtain the partisan summary for left political ideology. Similar prompt is used for the obtaining partisan summary for the right political ideology as well.

Topic Classification Task
You are given a 'document' and four labels, all derived from the same news event. The task is to determine the most topically relevant label to the document. Your goal is to assign the document to only one of the four labels. If the document is topically relevant to 'label1', please respond 'label1'. If the document is topically relevant to 'label2', please respond 'label2'. If the document is topically relevant to 'label3', please respond 'label3'. If the document is topically relevant to 'label4', please respond 'label4'. Strictly refrain from providing additional information. ### Input to analyze ### 'Document': {doc} 'Label1': {lab1} 'Label2': {lab2} 'Label3': {lab3} 'Label4': {lab4}

Table 16: Prompt template used for the topic classification task.

Partisan Classification Task
Given a segment of a 'news article' from a {ideology}-biased media source and two summaries derived from the same news event, your task is to perform binary classification by assigning the news article to one of the two summaries. Each summary has a set of talking points about the event. Each summary is also accompanied by metadata that includes frequently occurring actors, targets, sentiment on the target entities, and media frame associated with the context of the talking points. Your goal is to use the associated metadata to better determine if the provided news article segment has a viewpoint that is more similar to 'summary1' or 'summary2'. The response should strictly be 'summary1' when the 'news article' segment has a consistent viewpoint with 'summary1'; otherwise, it should be 'summary2' indicating the 'news article' has consistent viewpoint with 'summary2'. Refrain from providing any additional information. ### Input to analyze ### 'News article': {article} 'Summary1':{summ1} 'Summary1 Metadata':{summMetadata1} 'Summary2':{summ2} 'Summary2 Metadata':{summMetadata2}

Table 17: Prompt template used for the partisan classification task.

a smaller one (LLAMA3-8B), thereby instructing the smaller model to abstract away minor details and concentrate on event-level narratives.

Dataset. We select a portion of the dataset from our framework to fine-tune the model. Specifically, for each event, we consider the clusters from Section 3.2.2 and extract news articles linked to the talking points in each cluster, that are the most nearest to their respective ideology-specific views of that cluster. We take the top 25% of these articles to construct a dataset of 1100 examples, with the news articles as inputs and their corresponding partisan viewpoints as outputs. Each input-output pair includes a news article featuring two points of

view: one matching its ideological label and the other from an opposing ideology, which is crucial for training the LLM using preference optimization.

Training. We follow a two-stage training process. First, we start by instruction-tuning the model using SFT, where the news article is provided as the input and the viewpoints matching the ideological label of the article serves as the output. This enables the model to adjust its output probability distribution to focus on event-level discourse. For this, we train the model using PEFT and LORA adapter parameters for 1 epoch (with hyperparameters set to $r = 64, \alpha = 16, \textit{learning-rate} = 1e^{-4}$). Next, we train the model using Direct Preference Optimization (DPO) for 1 epoch, to align the model's generated outputs to prefer the viewpoints that match the ideology label of the article, while contrasting them with the viewpoints from the opposing ideology (hyperparameters: $r = 64, \alpha = 16, \textit{learning-rate} = 5e^{-5}$)). We fine-tuned the model (16-bit precision) using 8 Tesla V100 GPUs with 32GB VRAM each, and the training process took about 3 hours, for a batch size of 2 per device.

E Clustering the talking points

As described in 3.2.1, we cluster the initial talking point set to identify the prominent talking points. In this process, we merge redundant clusters and remove incoherent clusters. The details of this process is outlined in this section.

E.1 Merge Redundant Clusters

In order to merge redundant clusters, we devise a simple greedy algorithm. We consider pairwise combinations of cluster labels, and prompt the LLM to verify if the clusters can be merged based on the prompt template shown in 19.

We start by constructing a set of pairwise cluster labels $\mathcal{S} = \{(s_1, s_2), \cdots, (s_{n-1}, s_n)\}$. Precisely, for every cluster, we consider all possible pairwise combinations for the top-7 closest labels to that cluster in the embedding space. For each element in $\mathcal{S}$, we prompt LLM to infer if the pair of labels are discussing about the same aspect. If the aspects, say (s_1, s_2), are equivalent, then we merge these aspects, and update the set $\mathcal{S}$ by removing every element in the set that contains s_1 or s_2. In the second iteration, we construct a new set, $\mathcal{S}'$, that holds every combination of updated cluster labels, and repeat the previous step. We run the algorithm for two iterations or halt if there are no merges after the first iteration.

E.2 Remove Incoherent Clusters

We note that HDBSCAN algorithm provides us with an initial set of candidate clusters. For each candidate, we use the aspect associated with the cluster label to validate if the top-3 members that are closest to the cluster label in the embedding space are discussing the same high-level concept. We prompt the LLM using the prompt shown in 20 to remove incoherent clusters.

E.3 Talking Point Membership

After obtaining the cluster labels, which characterize the space of possible talking points. We consider each talking point from the set of all the talking points and assign the closest cluster label based on cosine similarity score. If this score is beyond a threshold value of 0.85, we assign the talking point to that cluster label. Otherwise, the it is discarded but retained in the unclustered pool of talking points.

E.4 Clustering-related Hyperparameters

Note that we are interested in identifying the dense regions in the embedding space associated with talking points, as these are the potential candidate topic indicators. Due to this, we choose HDBSCAN method as our clustering algorithm, which does not require any prior number of clusters. However, we are still required to tune a few hyperparameters in order to obtain a decent performance. We use a data-driven approach to estimate the best number of topics by maximizing the DBCV score(Moulavi et al. (2014)). We retain the default settings for *cluster_selection_method*, and *metric_parameters*, while we change the *min_cluster_size* and *min_samples* to get more sensible topics. This number is selected based on a grid search whose values are sensitive to the number of input talking points. Suppose |X| denote the number of talking points, then the grid parameters for HDBSCAN used in our method include $5, 7, 9, 0.01 * |X|, 0.02 * |X|, \cdots 0.04 * |X|$.

For our algorithm's talking point membership module, we choose a similarity threshold of 0.76 based on manually inspecting the prominent talking points, outputs for the cluster redundancy and removal of cluster incoherence operations for 3 events related to the issue - *Climate Change*.

Prompt to characterize a cluster
Given a set of news article excerpts taken from the same news event, the task is to analyze the articles with the intent to identify a high-level concept that captures the key aspect of discussion related to that event. The concept should be indicative of one of the main discussion angles related to the event, and not very specific to entities mentioned in the articles. The concept should be accompanied by a summary, which should not be a mere concatenation of articles.

Table 18: Prompt to characterize cluster candidate. We prompt the LLM in a two-shot setting.

Prompt to merge two clusters
Given two aspects from the same news event, you need to analyze them with the intent to understand if they are focusing on the same aspects of that event. You should compare the key emphasis of the aspects and their implications to decide if they are the 'same', or 'different'. Refrain from providing any additional explanations other than the label.

Table 19: Prompt to merge two clusters. We merge two clusters if their aspects are identical.

Prompt to remove inconsistent clusters
You are provided with a few news article excerpts and a key aspect of discussion, all of which are from the same news event. The task is to analyze if all of the provided news article excerpts are discussing the given key aspect. Respond with 'yes' if the central theme of discussion in each excerpt align in meaning with the key aspect, and 'no' if there is any variance, refraining from offering any additional explanation.

Table 20: Prompt to remove inconsistent clusters.

F Examples

Incoherent Example
Right Partisan Summary: R1: Title: Criticism of Biden's vaccine mandate as dictatorial Description: Dr. Fauci's characterization of President Biden's vaccine mandate as 'moderate' is challenged by right-leaning perspectives, labeling the mandate as dictatorial and overly lenient, emphasizing concerns about individual liberty being compromised. R2: Title: Opposition to perceived leniency in vaccine mandate Description: Right-leaning voices critique Dr. Fauci and President Biden for what they view as a lenient approach in the vaccine mandate, suggesting a need for stricter measures to ensure compliance and public health protection. R3: Title: Advocating for a stricter vaccination-only policy Description: Right-leaning perspectives highlight a preference for a stricter vaccination-only policy over the compromise of allowing testing as an alternative in President Biden's vaccine mandate, emphasizing the importance of prioritizing vaccination for public health benefits.
Topically relevant right talking points: 1. Title: President Biden's vaccine mandate is considered somewhat 'moderate' by Dr. Anthony Fauci Description: Dr. Fauci believes that President Biden's vaccine mandate strikes a balance between the need for vaccination and the concerns of those who do not want to get vaccinated but also do not want to lose their jobs. 2. Title: Dr. Fauci's stance on the vaccine mandate Description: Dr. Fauci is quoted as saying that the President's approach is somewhat moderate and a compromise, but the article disagrees, stating that Fauci's true stance on vaccine mandates is even less moderate. 3. Title: President Biden's vaccine mandates Description: Dr. Fauci describes President Biden's vaccine mandates as a moderate approach, allowing for testing as an alternative to vaccination for those who are unwilling to get vaccinated.
Corresponding news article summaries: 1. Dr. Fauci describes Biden's vaccine mandate as 'moderate' on CNN, emphasizing the need for options for those hesitant to get vaccinated. Republican governors criticize the mandate as an infringement on individual liberty and plan to challenge it in court. The mandate's impact on those previously infected with COVID sparks debate, with Fauci acknowledging the complexity of the issue. 2. A right-leaning article criticizes Dr. Fauci for supporting what they view as President Biden's overly lenient vaccine mandate approach, highlighting Fauci's perceived lack of stringency and labeling Biden's actions as dictatorial, while emphasizing the need to expose the true intentions of political figures like Fauci and Biden. 3. Dr. Fauci characterizes President Biden's vaccine mandate as moderate, emphasizing the option for testing as a compromise for those hesitant to get vaccinated, reflecting a right-leaning perspective on the level of stringency in vaccine mandates.

Table 21: Shows a negative example. The generated partisan summary is incorrect. This is primarily attributed to inconsistent news article summaries (2 and 3), and LLM's failure to identify cited information in the news article.