DiscoSum: Discourse-aware News Summarization

Anonymous ACL submission

Abstract

Recent advances in text summarization have predominantly leveraged large language models to generate concise summaries. However, language models often do not maintain long-term discourse structure, especially in news articles, where organizational flow significantly influences reader engagement. We introduce a novel approach to integrating discourse structure into summarization processes, focusing specifically on news articles across various media. We present a novel summarization dataset where news articles are summarized multiple times in different ways across different social media platforms (e.g. LinkedIn, Facebook, etc.). We develop a novel *news discourse schema* to describe summarization structures and a novel algorithm, **DiscoSum**, which employs beam search technique for structure-aware summarization, enabling the transformation of news stories to meet different stylistic and structural demands. Both human and automatic evaluation results demonstrate the efficacy of our approach in maintaining narrative fidelity and meeting structural requirements.

1 Introduction

In recent years, text summarization has seen remarkable advances, fueled by foundational Large Language Models that produce concise, context-rich overviews of lengthy documents (Li and Chaturvedi, 2024; Peper et al., 2024; Zhang et al., 2024). Yet despite these gains, current summarization approaches rarely account for a fundamental aspects of textual organization: discourse structure (Cohan et al., 2018).

Modern news organizations like the *the New York Times* increasingly publish news summaries in a variety of media (e.g. print newspapers, mobile apps, podcasts, and social media) each with distinct audience expectations and content formats (Kalsnes and Larsson, 2018; Ngoc, 2022). For instance, an outlet like The New York Times may produce a child-friendly podcast edition that uses simplified language and gentler framing, a condensed Instagram version with concise, visually engaging snippets, and a longer, more detailed write-up on LinkedIn or the newspaper's own website to cater to professional or academic readers. Transforming a single piece of news into multiple styles and lengths, while preserving its core narrative and emphasis, demands **nuanced control over discourse structure** (Shen et al., 2017; Hu et al., 2017).

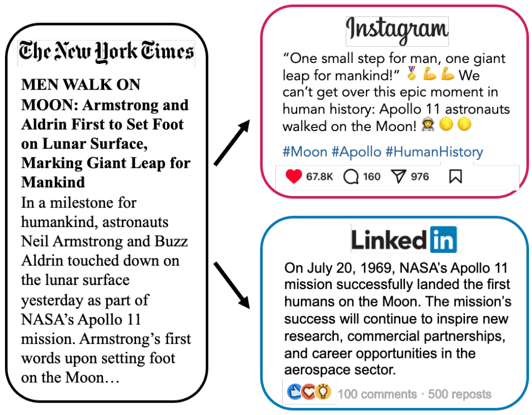

Figure 1: Comparative presentation of the Apollo 11 moon landing news across multiple platforms by The New York Times. This example showcases the diversity in content formatting and language adaptation for different audiences: a detailed traditional print article, a concise and visually-driven Instagram post, and a professionally oriented LinkedIn summary. Each platform reflects specific editorial strategies to engage its unique audience effectively, underlining the importance of discourse-aware news summarization.

Despite the growing interest in automated news summarization (See et al., 2017; Zhang et al., 2020; Beltagy et al., 2020; He et al., 2020; Zhao et al., 2022b), existing datasets approaches have overlooked this need[1]. To bridge these gaps, we propose a novel discourse-structure-aware summarization task that emphasizes the modeling of structural

[1]See Appendix C for a deeper comparison to Grusky et al. (2020).

discourse beyond surface-level summarization coherence or factual correctness.

First, we introduce **DiscoSum**: a **Disco**urse-aware News **Sum**marization dataset. DiscoSum represents the largest and most diverse collection of professionally-written cross-platform news summaries, comprising 20k news articles from 23 different news outlets across 10 countries, multiply paired with over 100k human-written summaries from 4 distinct platforms: Facebook, Instagram, Twitter and newsletters. Next, we develop a novel discourse schema to describe structural components of news summaries, consisting of five sentence-level discourse labels.Finally, we also propose a novel discourse-driven decoding method that employs a beam search technique to evaluate and select the optimal subsequent sentences for inclusion in summaries. We evaluate our method by developing both surface-level and structural metrics to assess the effectiveness of models in producing structure-aware summaries. Our human and automated evaluations confirm that our approach effectively maintains narrative fidelity and adheres to structural demands. In summary, we make the following contributions:

1. **New Task:** We introduce structure-aware summarization into the news domain.

2. **New Dataset:** We introduce a large-scale corpus of 20k news articles paired one-to-many with >100k different human-written summaries on Facebook, Twitter, Instagram and newsletters. We introduce a novel discourse schema for structural summarization.

3. **Benchmark Results:** We present baseline models and evaluations demonstrating the feasibility and potential of NLP systems for improving structure-aware news summarization.

2 Related Works

News Summarization. News summarization has been a key focus of natural language processing research (Barzilay and McKeown, 2005; Hong et al., 2014; Paulus et al., 2017; Goyal et al., 2022). Traditional methods often rely on extractive techniques, such as selecting "lead" sentences that approximate the news "lede" (Fabbri et al., 2019; Wang et al., 2020), but recent advancements in neural abstractive models have enabled more coherent and contextually rich summaries (Li and Chaturvedi, 2024; Peper et al., 2024; Zhang et al., 2024). Large-scale datasets, including those specifically curated for news articles, have further propelled model performance by providing diverse and representative training samples (Grusky et al., 2020; Chen et al., 2016). However, many of these approaches do not explicitly model the news article's inherent structure leading to summaries that, while fluent, may omit crucial structural components (Grenander et al., 2019; Zhao et al., 2022b).

Controllable Generation and Test-Time Alignment. Controllable generation has emerged as a promising way to ensure outputs satisfy certain style, tone, or length requirements (Yang et al., 2019; Yang and Klein, 2021; Zhao et al., 2022a). One notable area of research within controllable generation is *test-time alignment*, where models incorporate constraints or preferences at inference time to better conform to user or task-specific guidelines. Techniques such as prompt engineering or decoding-time gating have shown promise in guiding model outputs toward desired attributes (Meng et al., 2022; Huang et al., 2023; Liu et al., 2024). However, these methods often focus on surface-level constraints—like word count or style—and may not account for the deeper discourse structures characteristic of news articles.

Discourse-Aware Language Modeling. A growing body of work highlights the significance of discourse structures—such as identifying a document's truning points, sources, and concluding remarks—in improving text generation tasks (Zhai et al., 2003; Tian et al., 2024; Spangher et al., 2024c). Discourse-aware methods leverage theories like discourse elements (Spangher et al., 2022a) or journalistic guidelines (Spangher et al., 2022b) to parse and utilize the structural components of text during generation. While some efforts incorporate rhetorical roles or discourse parsing in domain-specific tasks (Wang and Cardie, 2013; Wang and Ling, 2016), their application to news articles is still nascent. By aligning with the natural organization of journalistic text, these methods show promise for generating summaries that both inform and engage, bridging the gap between factual coherence and audience-oriented design.

Our work differs from several prior attempts at structured summarization in other domains. The STRONG framework (Zhong and Litman, 2023) uses structure to parse legal documents to determine which elements to include in summaries, rather than controlling the generation structure itself. Similarly, work on dialogue summarization (Chen and Yang, 2023) employs structural controls such as entity tuples and dialogue act distributions but focuses on local coherence rather than global structure. Most similar to our approach is research on meta-review summarization (Shen et al., 2022), though it relies on hand-crafted and manually labeled articles with one-to-one article-to-summary mappings. Our work introduces methods for structured summarization without hand-labeling and allows for one-to-many mappings.

3 Task and Dataset

In this section, we describe the task formulation and evaluation metrics of structural summarization (§3.1). We introduce our proposed dataset including its composition and annotation process (§3.2).

3.1 Task Formulation

Let D denote the original news document, which can consist of multiple paragraphs or sentences. We define a desired sequence of discourse labels as $\mathbf{T} = (t_1, t_2, \ldots, t_n)$, where each t_i represents a discourse label (for instance, "contextual details," or "introductory elements," etc.) that the i-th sentence of the summary should fulfill. The objective is to generate a summary $\mathbf{S} = (s_1, s_2, \ldots, s_m)$, where each s_i is a sentence relevant to D and coherent. Note: In the *structured summarization* task we assume that the user supplies the target label sequence $\mathbf{T}$ *a priori*[2]. *Predicting* an optimal structure for new input is left for future work.

We employ a classification function $C(\cdot)$ that, given a sentence, predicts its discourse label. Let $\mathbf{L} = (l_1, l_2, \ldots, l_m)$ be the sequence of labels predicted by $C(s_i)$ for every sentence s_i in $\mathbf{S}$. We require $\mathbf{L}$ to align with $\mathbf{T}$ in order, so that $l_i = t_i$ for each position i. Although the most straightforward scenario sets $m = n$, such that the summary contains exactly n sentences, more flexible variants may allow for slight deviations while still ensuring that core positions match the targeted labels.

[2]This mirrors real newsroom workflows where social–media editors routinely apply pre–defined templates for different platforms. For example, commercial content–automation systems such as *Automated Insights* populate fixed headline and body layouts, and studies in discourse analysis show that canonical forms recur across news (Van Dijk, 1988; Dai et al., 2018) and even classical essay writing (De Montaigne, 1580).

Category	Count
# of Outlets	23
# of News Articles	20,811
# of Facebook Posts	18,275
# of Instagram Posts	66,030
# of Twitter Posts	8,977
# of Newsletters	10,506

Table 1: Overall counts of different categories.

Types	Counts
Overall	45,195
News Article → Tweet	12,516
News Article → Facebook Post	15,645
News Article → Instagram Post	7,738
News Article → Newsletter Post	9,296

Table 2: Statistics on the news article to summary graph, showing the number of edges between post types.

3.2 Dataset

We seek to construct a large, diverse dataset of news articles matched with multiple different summaries of each article, written by journalists, across different social media platforms and newsletters. We collect a list of 23 different major national and international news outlets[3] from 10 different countries (U.S., China, India, U.K., Germany, etc.), in order to capture a range of different discoure styles across different writing styles.

Social Media Collection We collect two years of social media posts on Twitter, Facebook and Instagram from each of the 23 news outlets. To do so, we build semi-automated scrolling agents that scroll down the feed of each news outlet's media page. We collect the full HTML of each post, including the text of each post as well as any linked urls. In total, we collect 8,977 Twitter posts, 18,275 Facebook posts, and 66,030 Instagram posts (see Tab. 2 for more details). In order to identify structural summaries, we further filter these posts down to posts that contain 50 or more characters. This eliminates around 30% of our data.

Newsletter Collection We select 7 newsletter brands published by news outlets,[4] specifically searching for those that make all past newsletters within each brand available online in archives. We

[3]The New York Times, The Wall Street Journal, Washington Post, AP News, BBC, Reuters, The Guardian, Bloomberg, Times of India, Le Monde, The New Zurich Times, El País, China Daily, Los Angeles Times, Chicago Tribune, The Boston Globe, USA Today, The Sydney Morning Herald, The Japan News, De Zeit

[4]*Axios* "The Finish Line"; *the New York Times*, "The Morning", *the LA Times*, "California Today"; *The Skimm*, "The Daily Skimm"; *The Daily Beast*, "Cheat Sheet"; *Semafor*, "Newsletters"; *CNN*, "Reliable Sources"

build scrapers to collect full HTML of each newsletter and collect 2 years worth of data, or over 20,000 newsletters (see Table Tab. 1 for details).

A newsletter often summarizes many news articles at the same time, yet our task is a single-document summarization task. Hence, we need to parse the text of each newsletter so that blocks of newsletter text are only corresponding to a single news article. This is *text segmentation with overlapping segments*, since newsletters might require larger continuous text segments that discuss multiple related topics. To accomplish this, we explored multiple prompts, comparing their effectiveness on a manually annotated subset of newsletters[5]. We selected a prompt configuration that instructs an LLM to (1) identify all news content links, (2) extract the surrounding text context for each link, (3) exclude boilerplate content, and (4) maintain the exact original text. To mitigate potential biases or hallucinations, we implemented a verification procedure where the largest extracted blocks are cross-checked against the LLM's own outputs in multiple iterations, with any inconsistencies flagged for manual review. This approach is supported by extensive research demonstrating LLM effectiveness for text segmentation tasks (Nayak; Zhao et al., 2024; Fan et al., 2024; Jiang et al., 2023). Manual inspection confirmed the LLM's capability in this task, with segmentation quality exceeding 95% accuracy in our audits across a randomly sampled set of 100 newsletters. In total, we generate 10,506 summaries from the newsletters we collect.

News Article Collection We collect a superset of news article URLs from all the social media posts and newsletters described above. Following Spangher et al. (2024a), we scrape Wayback Machine for the HTML of each news article. We use an LLM (GPT-4) to clean the HTML to extract a full, complete news article (we find that existing libraries[6] are insufficient). Our prompting strategy instructs the model to filter out non-news segments (e.g., login prompts, advertisements, and extraneous content), while retaining only article content.

News Article and Summary Matching For many social media posts, we have a URL in the post that gives us an explicit match; however, for others we do not (e.g. Instagram does not allow URLs in posts). To discover as many edges as possible, we decide to match *any* news article from *any* outlet with *any* social media post or newsletter summary. To do so, we employ a two-step rank-and-check method. Specifically, we first use SBERT (Reimers, 2019) to embed news articles and summaries; for each news article, we found the 10 closest summaries as candidates. Then, we use GPT-4 to perform a strict pairwise comparison for each candidate, returning only binary "yes" or "no" judgments on whether they describe the same news story, following the methodology validated in Spangher et al. (2024b)[7]. In manual audits, this matching step exceeds 95% accuracy, demonstrating the robustness of our multi-step procedure. Not only does this approach help us recover all summaries produced by a single news outlet for each article they publish, but we can see how *other* news outlets cover the same news event.

Dataset Splits For all experiments, we use a 70%/20%/10% train/validation/test (14k/4k/2k article-summary pairs) split of the DiscoSum dataset. This split is made at the article level to prevent leakage, so all summaries of the same article are kept within the same split.

4 Method

In this section, we first present how we construct a sentence-level labeler that predicts discourse labels (§4.1-§4.2). Based on this discourse labeler, we propose two strategies to generate summaries conforming to a target sequence of discourse labels $\mathbf{T}$: an edit-based approach (§4.3) and a beam search method (§4.4). The final goal is to ensure the predicted label sequence $\mathbf{L}$ aligns with $\mathbf{T}$.

4.1 Discourse Schema Generation

To formalize a notion of "structured" summaries, we seek to construct a low-dimensional, novel discourse schema to describe social media and newsletter summaries.

In contrast to prior work using manual analysis to develop schemas, typically based on O(10) examples[8], we sought to use an automated process to generate a schema. Inspired by Pham et al. (2024), we first ask an LLM to generate descriptive labels for the discourse role of each sentence in all of our summaries (O(100k) sentences). Then, we embed these labels using an SBERT embedding model

[5]Prompts shown in the Appendix.

[6]https://newspaper4k.readthedocs.io/en/latest/

[7]Authors found that LLMs could be used to verify cross-document event coreference with high performance.

[8]For example, Van Dijk (1988) builds their schema based on an analysis of 12 news articles.

(Reimers, 2019), and cluster these embeddings using k-means.

From this embedding process, we identify five distinct clusters that represent different narrative roles: Introductory Elements, Contextual Details, Event Narration, Source Attribution and Engagement Directive). See Tab. 4 for definitions of each discourse role. We confirm the validity of this schema by asking two professional journalists to assess the quality and ideate for missing role labels. The choice of specifically five discourse labels was informed by extensive experimentation. While alternative parameter choices (e.g., k=7, 13, or 23) were feasible in our clustering approach, we selected a 5-dimensional schema based on human evaluation trials that showed high inter-annotator agreement ($\kappa = 0.615$) for assessing the validity of these labels. Though a 5-dimensional schema may appear limited for capturing the full complexity of news discourse structures—particularly across cross-cultural or niche news scenarios—it provides a strong foundation for this pilot study in discourse-aware summarization.

4.2 Discourse Labeler

Following Spangher et al. (2021, 2022a), we construct a sentence-level labeler that serves as a context-aware classifier to assign discourse labels to sentences. The labeler was trained on the train split of DiscoSum. To verify the quality of the validation set, we had two expert annotators independently label a subset of 500 sentences. The trained labeler achieved a high accuracy rate of over 90% on the validation set, as shown in Fig. 4. This high level of accuracy is crucial for its role in the summarization process, where it is later used as a reward guidance mechanism to ensure that generated summaries adhere to the required discourse structure. The full confusion matrix in the appendix illustrates the labeler's strong performance across all five discourse categories, with the lowest per-category F1 score still exceeding 0.85.

4.3 Iterative Editing

Our first strategy approaches summary generation as an iterative refinement process. We begin by prompting the LLM to produce a complete initial summary, then repeatedly "edit" any sentences that do not fulfill their intended discourse labels. After the initial summary is generated, we use our labeler $C(\cdot)$ to identify which sentences carry the wrong labels. We then remove these "mismatched" sentences and generate new candidate sentences. Over several iterations, the summary gradually "evolves" to match the sequence $\mathbf{T}$.

Algorithm 1 Sentence-Level Discourse-driven Beam Search with Beam Size k

Require: Source text X, target label sequence $t_1, \ldots, t_N$, beam width k
Ensure: Best summary $S = \langle s_1, s_2, \ldots, s_N \rangle$

```
Initialize beam: 𝓑 ← {[]}                ▷ Start with an empty sequence
for i ← 1 to N do
    𝓑′ ← ∅
    for s ∈ 𝓑 do
        𝓒 ← LLM(s, X, k)                  ▷ Generate k candidate
        for c ∈ 𝓒 do
            s′ ← append(s, c)
            score ← C(s′, t_i)
            𝓑′ ← 𝓑′ ∪ {(s′, score)}
        end for
    end for
    𝓑 ← selectTopK(𝓑′, k)
end for
S ← argmax_{(s,score)∈𝓑} score
return S
```

By focusing only on individual problematic sentences, this approach preserves what is already correct in the summary. It can also adapt to complex label sequences without having to restart the entire generation each time a mismatch is found.

4.4 Sentence-Level Beam Search

In contrast to iteratively fixing errors, our second strategy constructs a label-compliant summary sentence by sentence from scratch in a beam search style (Lowerre, 1976).

We begin with an empty summary and consider one position at a time (e.g., first the sentence that should have the "introductory elements" label, then the sentence that should have the "contextual details" label, and so on). At each step i, the LLM generates several candidate sentences (forming a sentence-level "beam"), which are then evaluated by $C(\cdot)$. We choose the candidate that best matches the target label t_i. This sentence is appended to the current partial summary. By evaluating multiple options at each step and selecting the best match for the desired label, this approach ensures each summary sentence follows the intended label sequence. The detailed procedure is described as Alg. 1.

5 Experiments

In this section, we present our experimental setup (§5.1) and evaluation framework for structured summarization with target discourse labels (§5.2). We introduce baseline models and methods being benchamrked (§5.3). Next, we present empirical results (§5.4), human preference evaluation (§5.5) and the analysis on the impact of different beam sizes (§5.6).

5.1 Implementation Details

For vanilla generation, we sample the best output among 16 trials based on automated discourse labeler. In the Sentence-Level Beam Search, we employ BeamSize = 16. We fine-tuned the LLaMa-3-8B model using the PEFT method on the train split of **DiscoSum**. This fine-tuning approach reduced the validation loss significantly over 20 epochs. Key hyperparameters included a learning rate of 5e-05 and a multi-GPU distributed training setup across eight Nvidia 4090. For each generation in our experiments, we randomly generate a list of structural tags, to simulate the widest possible set of user inputs. This also prevented us from overfitting on commonly observed discourse structures.

5.2 Evaluation Protocols

Content Accuracy Evaluation. To quantify how the content accuracy of generated news summaries, we employ several metrics:

- **ROUGE-L.** (Lin, 2004) ROUGE-L, originally designed for summarization, measures the longest common subsequence of tokens between the generated summary and a reference summary.
- **FactCC.** (Kryscinski et al., 2020) FactCC is a model-based metric that classifies whether each generated sentence is factually consistent with the source document.
- **AlignScore.** AlignScore is a consistency metric that directly measures the factual correspondence between source and summary.

Structural Evaluation. To assess the alignment between the generated summary $\mathbf{S}$ and the expected discourse structure $\mathbf{T}$, we derive a predicted label sequence $\mathbf{L}$ from $\mathbf{S}$, formally

$$\mathbf{L} = \text{Labeler}(s_1, s_2, \ldots, s_n), s_i \in S$$

where Labeler represents either the human annotator or the automated model designed to identify discourse structures.

We employ three metrics to quantify the closeness of $\mathbf{L}$ to the target label sequence $\mathbf{T}$, which represents the ideal structural roles of sentences in the summary:

- **Longest Common Subsequence (LCS).** LCS measures the length of the longest subsequence common to $\mathbf{L}$ and $\mathbf{T}$. A higher LCS value indicates that the predicted labels closely preserve the intended label order.
- **Match Score.** The Match Score assesses the number of exact position-wise matches between $\mathbf{L}$ and $\mathbf{T}$. This metric reflects the precision in predicting each label at its correct position in the sequence.
- **Levenshtein Distance.** (Levenshtein, 1965) This metric calculates the minimum number of single-element edits (insertions, deletions, or substitutions) required to transform $\mathbf{L}$ into $\mathbf{T}$. A lower Levenshtein Distance indicates a higher degree of sequence similarity.

Given the potential high cost of human evaluation, we provide protocols for both automated and human assessments:

Human Evaluation. We worked with two human annotators to manually assess the discourse structure of each sentence in the generated summaries. In this study, we ask annotators to evaluate 100 summaries for each model.

5.3 Baselines

To evaluate the effectiveness of our proposed approach, we benchmark it against a range of baseline models that vary in architecture, training paradigms, and optimization goals. These models include both proprietary systems and open-source alternatives, providing a comprehensive overview of current state-of-the-art capabilities in text summarization and related tasks.

Close-source LLMs. These models, such as DeepSeek-V3 [9], Claude-3-5-sonnet [10], and GPT-4o [11], are included primarily to help us gauge how well our approach performs in comparison to leading-edge technology, even if these models are not the primary focus of our evaluation.

[9] https://api-docs.deepseek.com/news/news1226
[10] https://www.anthropic.com/claude/sonnet
[11] https://openai.com/index/hello-gpt-4o/

Open-Source LLMs. Models like Qwen-2.5 and various configurations of LLaMa-3-8B represent more accessible options that are widely used in academic research. Each variant of LLaMa-3-8B—whether it be the vanilla version, edit-based modifications, or fine-tuned iterations—serves to illustrate different potential improvements and trade-offs within the open-source framework.

5.4 Main Results

Content Accuracy Evaluation. Table 3 shows both surface-level and structural evaluations for a variety of models. Despite fluctuations in ROUGE-L, FactCC, and AlignScore across different systems, our approach—specifically the beam search variant of LLaMa-3-8B—maintains competitive performance in surface-level metrics. Notably, our beam search method achieves the highest AlignScore (0.3890), demonstrating superior factual consistency with source documents compared to both proprietary and other open-source models. This is particularly significant as it shows that structural improvements can be achieved without sacrificing—and in fact can enhance—factual alignment with source content. We also include the reasoning–centric model *O1*, which outperforms GPT-4o on several metrics yet still lags behind our LLaMa-3-8B beam–search variant.

Structural Evaluation. Significantly, our approach excels in both automatic and manual structural evaluations, where it demonstrates notable enhancements over both open-source baselines and the more sophisticated proprietary models. The beam search variant of LLaMa-3-8B consistently aligns more closely with the designated discourse label sequences, evidenced by its superior Match Score and reduced Levenshtein Distance. This enhancement in structural alignment underscores the model's ability to adhere rigorously to specified rhetorical structures without significant loss in surface-level accuracy. By achieving an effective balance between textual overlap and structural fidelity, our method significantly enhances the controllability and coherence of generated text.

Performances of Edit-based and Finetuned Methods. The edit-based method demonstrates a promising capability in enhancing the structural alignment of generated summaries with the desired discourse labels, as evidenced by its strong performance in structural evaluations. However, this structural fidelity comes at a cost to the content accuracy and fluency, where the ROUGE-L scores considerably lower than other methods. This decline indicates that while the edit-based approach effectively molds the structure of the summaries, it may deviate significantly from the original text's semantic and syntactic properties.

The finetuned variant of the LLaMa-3-8B model, on the other hand, shows a less impressive adaptation to the task. Despite the potential for finetuning to tailor model behavior closely to specific datasets or task requirements, the observed performance metrics suggest a failure to capture the deeper, structural nuances necessary for this specific discourse-driven summarization task. The low scores imply that mere finetuning may be insufficient for tasks that require a deep understanding and transformation of text according to complex labeling schemes. This underperformance highlights the need for more advanced approaches to our task.

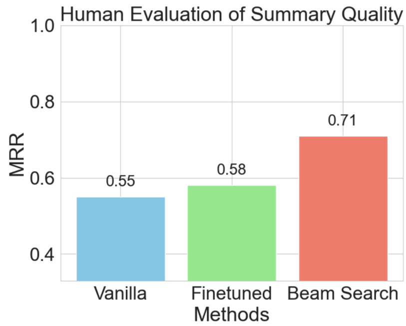

Figure 2: Mean Reciprocal Rank (MRR) scores from human preference evaluations of summary quality across three methods: Vanilla LLaMa-3-8B, Fine-tuned LLaMa-3-8B, and Beam Search LLaMa-3-8B.

5.5 Human Evaluation of Summary Quality

We recruited two annotators to ranked the summaries based on content accuracy and structural adherence for three summary generation methods—Vanilla LLaMA-3-8B, its fine-tuned counterpart, and our beam search method. Our results, depicted in Figure 2, demonstrate a significant superiority of the beam search method, achieving a mean reciprocal rank (MRR) of 0.71, compared to 0.55 and 0.58 for the Vanilla and fine-tuned methods, respectively.

5.6 The Impact of Beam Size

Our analysis incorporated a range of beam sizes from 2 to 16. As the beam size increases, we observe an overall improvement in the LCS scores, indicating enhanced alignment with the target discourse structure. Conversely, the Levenshtein Dis-

	Content Accuracy			Auto Struct.			Human Struct.		
Models	**R-L (%) ↑**	**FactCC ↑**	**AlignScore ↑**	**MS ↑**	**Lev ↓**	**LCS ↑**	**MS ↑**	**Lev ↓**	**LCS ↑**
Prop. Models									
DeepSeek-V3	47.15	0.47	0.3886	0.26	0.64	0.65	0.24	0.65	0.65
Claude	34.30	**0.70**	0.3882	0.25	0.68	0.64	0.20	0.49	0.75
GPT-4o	29.51	0.63	0.3884	0.11	0.80	0.62	0.15	0.58	0.68
O1	44.65	0.50	-	0.28	0.66	0.54	-	-	-
Open-sourced Models									
Qwen-2.5	40.82	0.58	0.3888	0.24	0.66	0.65	0.15	0.52	0.64
LLaMa-3-8B	**47.18**	0.50	0.3496	0.21	0.77	0.36	0.24	0.49	0.65
– Finetuned	22.01	0.61	0.3495	0.14	0.77	0.45	0.18	0.55	0.72
– Edit-based	15.28	0.59	-	0.51	0.48	0.56	0.24	0.65	0.36
– Beam Search	42.98	0.64	**0.3890**	**0.72**	**0.32**	**0.68**	**0.55**	**0.17**	**0.87**

Table 3: Comparison of models on various metrics. Metrics are categorized into content accuracy and structural assessments, both automated and human-annotated. The metrics include ROUGE-L (%), FactCC, AlignScore (for factual consistency), Match Score (MS), Levenshtein Distance (Lev), and Longest Common Subsequence (LCS). ↑ for higher is better and ↓ for lower is better. Boldfaced numbers highlight the best performance, while underscored numbers denote notable but secondary performances in each category.

tance, which measures the edit distance necessary to align the predicted sequence with the target, exhibits a general decrease as the beam size increases, suggesting that larger beam sizes improve structural alignment.

The observed trends open several avenues for future research. One potential area is the exploration of adaptive beam sizes that could dynamically adjust based on the complexity of the text or the specific requirements of the discourse structure at different points in a document. Additionally, while beam search techniques enhance the quality and relevance of summaries during the inference time, integrating these high-quality summaries during training could potentially elevate the model's overall performance. Future research could look into harnessing these refined outputs to boost the training process.

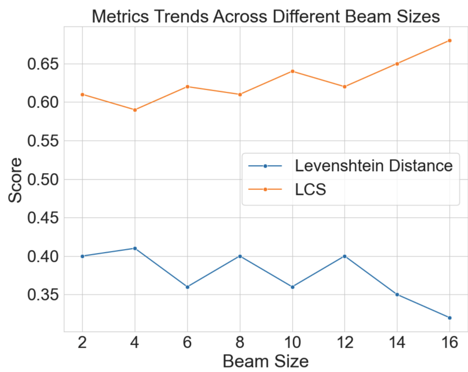

Figure 3: Levenshtein Distance and Longest Common Subsequence (LCS) scores as a function of beam size in structured summarization. The graph shows a general decrease in Levenshtein Distance and a gradual increase in LCS scores, indicating improved structural alignment with larger beam sizes.

6 Conclusion

In this study, we introduced a structural summarization approach that integrates discourse organization into the summarization of news articles, emphasizing narrative fidelity and structural alignment. Our novel dataset, DiscoSum, and evaluation metrics underscore the effectiveness of our methods, particularly the beam search technique, which ensures summaries are both contextually relevant and structurally precise. The results demonstrate significant improvements over traditional methods, suggesting that our approach can enhance automated news summarization across diverse media platforms.

Our contributions highlight unexplored research problems in the field of news summarization. Our **DiscoSum** dataset and corresponding evaluation metrics set the foundation for further exploration into how discourse elements can be systematically incorporated into summarization models. This shift towards a deeper understanding of discourse structures not only challenges existing models but also opens pathways for more sophisticated approaches to news narrative reconstruction. By emphasizing structuring over surface-level coherence, we invite the research community to explore novel methodologies that could change how news content is summarized across diverse media landscapes (Caswell and Dörr, 2018; Spangher et al., 2022b; Caswell, 2024; Welsh et al., 2024).

Limitations

Focus of the Study. Although we measure content accuracy using standard metrics (e.g., FactCC, ROUGE-L) and acknowledge its importance, our primary goal is to ensure structural alignment with discourse labels rather than to optimize factual correctness. Consequently, improvements in factual precision or content coverage are incidental rather than intentional. Future work could investigate techniques that integrate more robust fact-checking and retrieval-augmented generation to complement structural fidelity, particularly in applications where factual accuracy is critical.

Trade-offs in Decoding Efficiency. While our beam search method significantly improves structural adherence, it can be more computationally expensive compared to simpler generation techniques. This overhead may pose a challenge for real-time applications or large-scale deployment. Future research could explore adaptive beam strategies or hybrid methods that balance decoding speed with the need for strict discourse control.

Potential Data Biases. Our data collection methodology involves LLMs for several critical tasks, including HTML cleaning, newsletter segmentation, and article-summary matching. While we have taken extensive steps to validate these processes, these models may introduce biases that affect the dataset's composition and the resulting schema. To mitigate this concern, we collected a diverse dataset spanning 23 major news outlets from 10 different countries across 4 different distribution methods, which helps balance potential biases across different writing styles and outlet preferences.

Additionally, while our discourse schema is intentionally coarse-grained to enhance generalizability, we acknowledge that biases in structures can still occur. Although our primary focus was on structural rather than lexical aspects, entity or gender biases identified in prior work (Spangher et al., 2024a) could potentially percolate to structural patterns. The size and diversity of our dataset help mitigate these concerns, but future work should explore the relationship between lexical biases and discourse structures, particularly for applications that require cross-cultural or domain-specific adaptations.

References

Regina Barzilay and Kathleen R McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

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

David Caswell. 2024. Telling every story: Characteristics of systematic reporting. In *Journalism and Reporting Synergistic Effects of Climate Change*, pages 266–283. Routledge.

David Caswell and Konstantin Dörr. 2018. Automated journalism 2.0: Event-driven narratives: From simple descriptions to real stories. *Journalism practice*, 12(4):477–496.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Association for Computational Linguistics (ACL)*.

Jiaao Chen and Diyi Yang. 2023. Controllable conversation generation with conversation structures via diffusion models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7238–7251.

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

Zeyu Dai, Himanshu Taneja, and Ruihong Huang. 2018. Fine-grained structure-based news genre categorization. In *Proceedings of the Workshop Events and Stories in the News*, pages 17–23.

Michel De Montaigne. 1580. *Essays*. Self-published.

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.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2024. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16998–17010.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. 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 6019–6024, Hong Kong, China. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2020. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *Preprint*, arXiv:1804.11283.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *arXiv preprint arXiv:2012.04281*.

Kai Hong, John M Conroy, Benoit Favre, Alex Kulesza, Hui Lin, Ani Nenkova, et al. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *LREC*, pages 1608–1616.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Tenghao Huang, Ehsan Qasemi, Bangzheng Li, He Wang, Faeze Brahman, Muhao Chen, and Snigdha Chaturvedi. 2023. Affective and dynamic beam search for story generation. *ArXiv*, abs/2310.15079.

Feng Jiang, Weihao Liu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Haizhou Li. 2023. Advancing topic segmentation and outline generation in chinese texts: The paragraph-level topic representation, corpus, and benchmark. *arXiv preprint arXiv:2305.14790*.

Bente Kalsnes and Anders Olof Larsson. 2018. Understanding news sharing across social media: Detailing distribution on facebook and twitter. *Journalism studies*, 19(11):1669–1688.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Haoyuan Li and Snigdha Chaturvedi. 2024. Rationale-based 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 8274–8292, Mexico City, Mexico. Association for Computational Linguistics.

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.

Qin Liu, Fei Wang, Nan Xu, Tianyi Yan, Tao Meng, and Muhao Chen. 2024. Monotonic paraphrasing improves generalization of language model prompting. *ArXiv*, abs/2403.16038.

Bruce T. Lowerre. 1976. The harpy speech recognition system.

Tao Meng, Sidi Lu, Nanyun Peng, and Kai-Wei Chang. 2022. Controllable text generation with neurally-decomposed oracle. *ArXiv*, abs/2205.14219.

Kota Shamanth Ramanath Nayak. Does chatgpt measure up to discourse unit segmentation? a comparative analysis utilizing zero-shot custom prompts.

Nguyen Minh Ngoc. 2022. Journalism and social media: The transformation of journalism in the age of social media and online news. *European Journal of Social Sciences Studies*, 7(6).

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *Preprint*, arXiv:1705.04304.

Jospeph J. Peper, Wenzhao Qiu, and Lu Wang. 2024. Pelms: Pre-training for effective low-shot multi-document summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. Topicgpt: A prompt-based topic modeling framework. 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 2956–2984.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *Preprint*, arXiv:1704.04368.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. Mred: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2521–2535.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Preprint*, arXiv:1705.09655.

Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 498–517.

Alexander Spangher, Yao Ming, Xinyu Hua, and Nanyun Peng. 2022a. Sequentially controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6848–6866, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexander Spangher, Nanyun Peng, Sebastian Gehrmann, and Mark Dredze. 2024a. Do llms plan like human writers? comparing journalist coverage of press releases with llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21814–21828.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022b. NewsEdits: A news article revision dataset and a novel document-level reasoning challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–157, Seattle, United States. Association for Computational Linguistics.

Alexander Spangher, Serdar Tumgoren, Ben Welsh, Nanyun Peng, Emilio Ferrara, and Jonathan May. 2024b. Tracking the newsworthiness of public documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14150–14168, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Spangher, James Youn, Matt DeButts, Nanyun Peng, Emilio Ferrara, and Jonathan May. 2024c. Explaining mixtures of sources in news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15837–15859, Miami, Florida, USA. Association for Computational Linguistics.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Teun A Van Dijk. 1988. *News as discourse*. Routledge.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6209–6219, Online. Association for Computational Linguistics.

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405, Sofia, Bulgaria. Association for Computational Linguistics.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

Ben Welsh, Naitian Zhou, Arda Kaz, Michael Vu, and Alexander Spangher. 2024. Newshomepages: Homepage layouts capture information prioritization decisions. *Preprint*, arXiv:2501.00004.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Shuai Yang, Zhangyang Wang, Zhaowen Wang, Ning Xu, Jiaying Liu, and Zongming Guo. 2019. Controllable artistic text style transfer via shape-matching gan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4442–4451.

ChengXiang Zhai, William W. Cohen, and John D. Lafferty. 2003. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

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

Chao Zhao, Faeze Brahman, Tenghao Huang, and Snigdha Chaturvedi. 2022a. Revisiting generative commonsense reasoning: A pre-ordering approach. *ArXiv*, abs/2205.13183.

Chao Zhao, Tenghao Huang, Somnath Basu Roy Chowdhury, Muthu Kumar Chandrasekaran, Kathleen McKeown, and Snigdha Chaturvedi. 2022b. Read top news first: A document reordering approach for multi-document news summarization.

In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 613–621, Dublin, Ireland. Association for Computational Linguistics.

Jihao Zhao, Zhiyuan Ji, Yuchen Feng, Pengnian Qi, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Meta-chunking: Learning efficient text segmentation via logical perception. *arXiv preprint arXiv:2410.12788*.

Yang Zhong and Diane Litman. 2023. Strong–structure controllable legal opinion summary generation. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 431–448.

A Confusion Matrix of Discourse Labeler

In this section, we present the confusion matrix of our trained discourse Labeler. The overall accuracy is 90.90% and F-1 score is 0.9087. The results demonstrate the robustness our trained discourse labeler.

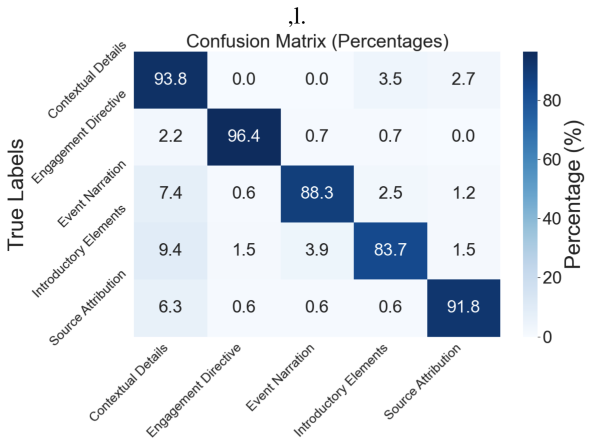

Figure 4: Confusion Matrix of Discourse Labeler.

B Discourse Schema Definition

In this section, we present definitions of the discourse schema.

C Comparison to Newsroom

The NEWSROOM dataset (Grusky et al., 2020) is a widely used resource for summarization research that, like our work, contains news articles and their summaries. However, our DiscoSum dataset differs fundamentally in its collection methodology, content richness, and research focus.

C.1 Collection Methodology

NEWSROOM's collection mechanism extracts summaries from HTML metadata, specifically the `<meta property="description">...</meta>` tags embedded in article URLs. This approach efficiently collects a large volume of summaries but is limited to a single summary per article that was intended for search engine or link preview contexts.

In contrast, DiscoSum collects actual posts written by journalists for specific social media platforms and newsletters. Modern newsrooms typically employ dedicated social media teams that craft platform-specific content, resulting in multiple distinct summaries of the same article across different channels. These summaries are rarely exposed in the article's HTML metadata, as they are written directly into each platform's publishing interface.

C.2 Content Comparison

To illustrate this difference, we present a case study of a New York Times article about uniquely shaped Yankees baseball bats:

As demonstrated in Table 5, the meta description (collected by NEWSROOM) is brief and focuses narrowly on the analyst's career move. In contrast, the social media posts (collected by DiscoSum) provide richer information about the story's core elements—the innovative bat design, the physics behind it, and quotes from the creator—with varying levels of detail across platforms.

C.3 Research Value

DiscoSum offers several advantages for summarization research:

1. **Multiple reference summaries per article:** DiscoSum provides multiple professionally written summaries for each article, spanning different platforms and formats.

2. **Platform-specific structural patterns:** The dataset captures how the same content is adapted for different platforms (Twitter, Facebook, Instagram, newsletters), revealing platform-specific structural patterns.

3. **Real-world audience targeting:** The summaries in DiscoSum represent actual content seen by users, written by professional journalists with specific audience and platform considerations in mind.

4. **Discourse structure analysis:** By annotating these varied summaries with discourse labels, DiscoSum enables research into how narrative structures adapt across platforms.

Label	Definition
Introductory Elements	Sets the stage for the summary by introducing the main topic, themes, or key points that will be covered.
Contextual Details	Provides additional background and setting information to help understand the main events or topics being summarized.
Engagement Directive	Directs the reader's attention or actions through calls to action, questions, or direct addresses to engage them with the content.
Event Narration	Describes specific events or occurrences in a narrative form, detailing what happened in a sequential or explanatory manner.
Source Attribution	Cites the origins of the information, giving credit to sources or clarifying the basis of the claims made in the summary.

Table 4: Discourse Label Definitions for Structured Summarization

Source	**Content**
Article URL	https://www.nytimes.com/athletic/6241862/2025/03/30/yankees-bats-aaron-leanhardt-marlins/
Meta Description (NEWSROOM)	"Aaron Leanhardt was the Yankees lead analyst in 2024 before joining the Marlins coaching staff this offseason."
Facebook Post (DiscoSum)	"The New York Yankees' uniquely shaped bats have caught the attention of many and are the result of two years of research and experimentation."
Twitter Post (DiscoSum)	"From @TheAthletic: The New York Yankees' uniquely shaped bats have caught the attention of many and are the result of two years of research and experimentation. Meet the former MIT physicist behind the 'torpedo' bats."
Instagram Post (DiscoSum)	"The New York Yankees' uniquely shaped bat is the result of two years of research and experimentation with a former MIT physicist-turned-coach at the helm. Aaron Leanhardt, the brains behind the 'torpedo bats' making headlines, says the idea behind his innovation was simple — redistribute the weight of the bat to where it matters. The bats have been around for more than just this season. Players used them in 2024. But after last weekend's laser show in the Bronx, they have broken into the mainstream. 'Ultimately, it just takes people asking the right questions and being willing to be forward-thinking,' Leanhardt says."

Table 5: Comparison of content collected by NEWSROOM versus DiscoSum for the same New York Times article. Notice how different the formats are for different social media platforms.

While NEWSROOM has been invaluable for general summarization research, DiscoSum specifically enables the study of discourse-aware summarization strategies that can adapt to different platforms and structural requirements—a capability increasingly important as news consumption fragments across diverse digital channels.

D Prompts

D.1 Leaf Node Prompt

This prompt is used to analyze and categorize discourse roles in news articles by summarizing common patterns across a set of annotated sentences. It generates concise labels that capture the essential discourse function of a group of sentences.

```
You are a helpful assistant. I will give
    you a large set of notes about
    sentences in news articles that I
    wrote down.

Here are the notes:

{labels}

Please summarize them, focusing on the
    common discourse role each sentence
    plays, based on the notes. Ignore
    the topic.
Summarize them with a single, specific
    label for the entire group, being
    sure to concisely capture what they
    are about.
Make the label 2-3 words, max. Be
    descriptive but not too broad.
    Please return just one label and one
     description.
Make it in this format: ```"Label":
    Description```
```

D.2 Middle Tree Prompt

This prompt is designed for hierarchical categorization of writing elements. It helps create mid-level labels that group similar discourse roles together, focusing on common functional aspects while ignoring specific topics.

```
You are a helpful assistant. I will give
     you a notes about different writing
     elements.

Here are the notes:

{labels}

Please summarize them with one label,
    focusing on the common discourse
    role each element plays, ignoring
    the topic.
Summarize them with a single, specific
    label for the entire group,
    concisely capturing what they are
    about.
Make the label 2-3 words, max. Be
    descriptive but not too broad.
    Please return just one label and one
     description.
Make it in this format: ‘‘‘"Label":
    Description‘‘‘
```

D.3 Few-Shot Example Selection Prompt

This prompt is used to identify representative examples for each discourse label. It selects diverse, high-quality examples that best illustrate a particular label, which can later be used for few-shot learning or annotation guidelines.

```
I am trying to find good examples to use
     for demonstrating a label.
Here is the label: {label}. The
    definition for the label is: {
    definition}.

Here are a large set of examples I have,
     alone with notes for each one:
[Start Examples]
{examples}
[End Examples]

Some examples are bad. Please choose 4
    examples that best represent this
    label. Try to pick diverse ones.
Return the examples and the notes, and
    copy them fully.
Return as a json. Be careful to format
    the quotes correctly.
```

D.4 Definitions Prompt

This prompt assigns predefined discourse role labels to sentences within social media posts. It uses contextual information from both the full post and specific annotations to match sentences with the most appropriate discourse label from a controlled vocabulary.

```
I will give you a social media post and
    a single sentence from that post.
Your goal is to assign a label to that
    sentence with a general discourse
    role that best describes it’s
    purpose in the overall script.
Each sentence also includes some notes I
     took about the very specific
    discourse role it plays, you can use
     them if it’s helpful.

Choose from this list:
{discourse_labels}

Do NOT return any labels NOT in that
    list. Here are some shortened
    examples:
‘‘‘{examples}‘‘‘

Now it’s your turn. Here is a social
    media post:
‘‘‘{full_document}‘‘‘

What discourse role is this sentence in
    it serving?
Sentence: ‘‘‘{sentence}‘‘‘
Notes: ‘‘‘{notes}‘‘‘
Answer:
```

D.5 Newsletter Processing Prompt

This prompt extracts and organizes news content from newsletters. It identifies and separates text blocks associated with specific links, focusing on meaningful news content while filtering out boilerplate text.

```
Look at the clean HTML of this
    newsletter.

Please separate the blocks of text into
    news content corresponding to each
    individual link.
This includes all the context
    surrounding the links.
Exclude links that do not pertain to
    news content.
The same text can be included in
    different chunks if it is relevant
    to a link.
Try to include all text in at least one
    chunk.
If a line doesn’t end with a period,
    please add one.
Do not change the text otherwise, in any
     way.
Ignore text that is boilerplate and not
    related to news content.

Return a python dictionary mapping the
    link to each chunk of text. Don’t
    return anything else. Copy the text
    exactly.

    ‘‘‘{html}‘‘‘
```

E Post Examples

In this section, we present posts examples with discourse labels.

Label	sentence
Introductory Elements	Boston's streets are changing.
Contextual Details	A growing number of them have bike lanes meant to protect bicyclists, slow down drivers, reduce the risk of crashes, and ultimately get more people to feel comfortable biking
Introductory Elements	The city is aiming to expand the bike lane network so that half of residents live within a 3-minute walk of a safe and connected bike route by the end of next year.
Contextual Details	The theory is that if there is a safe path for biking, more people will take it, in turn reducing climate change-causing emissions, traffic deaths, and mind-numbing congestion.
Engagement Directive	But challenges remain.
Introductory Elements	Many projects face vocal opposition to ceding valuable street real estate to bikes.
Introductory Elements	And other issues, such as the prevalence of large trucks, and lingering gaps in the bike network, make biking more dangerous than most would like.

Table 6: An example of Instagram post with sentence-level labels

Label	sentence
Introductory Elements	Global upheaval has once again increased America's geopolitical importance.
Event Narration	This years election campaign will shape the direction of U.S. policy.
Contextual Details	It is thus being closely watched around the world.

Table 7: An example of facebook post with sentence-level labels

Label	sentence
Introductory Elements	Logan Edra, a 21-year-old American B-Girl, said the Olympics could provide young girls with a vision of the future.
Source Attribution	"Any type of representation is going to help people see what is possible."

Table 8: An example of twitter post with sentence-level labels

Label	sentence
Event Narration	Disney began laying off thousands of staff members, its second round of layoffs, to save $5.5 billion in costs and cut 7,000 jobs.
Contextual Details	Employees at ESPN, Disney Entertainment, Disney Parks, and Experiences and Products will also be affected.
Engagement Directive	A third round of layoffs is expected before summer.
Introductory Elements	Meanwhile, Insider's employees went on strike after about 10% of its staff was laid off.
Contextual Details	Staffing cuts have also affected Buzzfeed, NPR, and other news organizations.

Table 9: An example of newslttter with sentence-level labels

Original News Article	
There's overwhelming support for an age limit on the president and Congress, but it won't happen anytime soon. Polling by Insider and Morning Consult indicates that three in four Americans favor an age limit for members of Congress. More than four in 10 viewed the ages of political leaders as a "major" problem. Some lawmakers are concerned that politics has become a gerontocracy, and the country's oldest politicians have not just the incentive to stick around but the power to remain entrenched. The seniority system, a campaign-finance system that favors incumbents, gerrymandered congressional districts, and the ability to continue to seek office indefinitely are mechanisms that members of Congress are most eager to discuss. Age limits are not on the table, and Congress' youngest and oldest lawmakers seem to agree. The country's power class is also aging across the board; the current Congress is the oldest in the nation's history. At the start of the latest Congress, in January 2021, baby boomers controlled roughly 56% of the seats in both chambers despite representing approximately just 21% of the US population. Some lawmakers reflect on an underlying system that prioritizes tenure and seniority over other potential measures of political acumen, generating a political leadership that skews older. Term limits offer another potential solution, though that too would require going through the arduous constitutional-amendment process."	
Target Labels: Engagement Directive, Contextual Details, Introductory Elements, Introductory Elements, Event Narration	
Engagement Directive	Many are calling for a constitutional amendment to set a mandatory retirement age for lawmakers, citing the need for fresh perspectives and a more diverse range of experiences in Washington.
Contextual Details	Additionally, the proposed age limit aims to address concerns that the current age distribution in Congress may hinder its ability to effectively represent the diverse needs and values of the younger generation.
Introductory Elements	Despite the widespread support, opponents argue that imposing an age limit would violate the constitutional principle of equal protection, as it would unfairly target a specific group based on age.
Introductory Elements	Despite the ongoing debate, polling suggests that three in four Americans support an age limit for members of Congress, with over four in 10 viewing the ages of political leaders as a major problem.
Event Narration	A recent poll has found that three in four Americans favor an age limit for members of Congress, with over four in 10 viewing the ages of political leaders as a "major" problem.

Table 10: Example generation #1

Original News Article
Are robot waiters the future? Some restaurants think so. MADISON HEIGHTS, Mich. (AP) ˇ2014 You may have already seen them in restaurants: waist-high machines that can greet guests, lead them to their tables, deliver food and drinks and ferry dirty dishes to the kitchen. Some have cat-like faces and even purr when you scratch their heads. But are robot waiters the future? Itˇ2019s a question the restaurant industry is increasingly trying to answer. Many think robot waiters are the solution to the industryˇ2019s labor shortages. Sales of them have been growing rapidly in recent years, with tens of thousands now gliding through dining rooms worldwide. Dennis Reynolds, dean of the Hilton College of Global Hospitality Leadership at the University of Houston, says, There's no doubt in my mind that this is where the world is going. The school's restaurant began using a robot in December, and Reynolds says it has eased the workload for human staff and made service more efficient. However, others say robot waiters aren't much more than a gimmick that have a long way to go before they can replace humans. They can't take orders, and many restaurants have steps, outdoor patios and other physical challenges they can't adapt to. Redwood City, California-based Bear Robotics introduced its Servi robot in 2021 and expects to have 10,000 deployed by the end of this year in 44 U.S. states and overseas. Shenzen, China-based Pudu Robotics, which was founded in 2016, has deployed more than 56,000 robots worldwide. Li Zhai, owner of Noodle Topia in Madison Heights, Michigan, says the robots give human servers more time to mingle with customers, which increases tips. And customers often post videos of the robots on social media that entice others to visit. But not all chains have had success with robots. Chili's introduced a robot server named Rita in 2020 and expanded the test to 61 U.S. restaurants before abruptly halting it last August. The chain found that Rita moved too slowly and got in the way of human servers. Eventually, experts expect that a certain percentage of restaurants maybe 30% will continue to have human servers and be considered more luxurious, while the rest will lean more heavily on robots in the kitchen and in dining rooms.
Target Labels: Event Narration, Contextual Details, Event Narration, Engagement Directive, Event Narration

Event Narration	Robot waiters are here to stay, and their presence is redefining the future of the restaurant industry.
Contextual Details	The innovative technology, first introduced at a Michigan restaurant, has been gaining popularity worldwide, as they bring efficiency, convenience, and a touch of novelty to the dining experience.
Event Narration	Robot waiters are increasingly being adopted by restaurants worldwide, offering benefits such as reduced labor costs, improved accuracy, and enhanced customer engagement.
Engagement Directive	As the technology continues to evolve, we can expect to see robot waiters becoming more sophisticated, with advanced navigation systems and the ability to learn and adapt to individual customers' preferences.
Event Narration	With the potential to transform the dining experience, robot waiters are being hailed as a game-changer in the industry, enabling restaurants to adapt to changing consumer preferences and improve overall efficiency, as well as providing a unique and entertaining experience for customers.

Table 11: Example generation #2

Original News Article	
D̈on't fool yourself̈: Billions more needed to protect tropical forests, warns new report. At least $130bn a year is needed to protect the most at-risk areas of tropical forest by the end of the decade, alongside reductions in beef and dairy consumption and government bans on deforestation, a thinktank has warned. Currently, finance to protect forests averages between $2bn and $3bn a year. The report estimates that eliminating the economic incentive to destroy forests for cattle ranching, agriculture and other uses would cost at least $130bn a year. The money could come from carbon markets, wealthy governments and philanthropists, but there must also be urgent actions such as a ban on clearing forests, developing businesses that rely on standing forests and reducing demand for commodities linked to deforestation, such as palm oil, soya, beef and cocoa. Lord Turner, a former head of the CBI and ex-chair of the UK government's Committee on Climate Change, warned that governments should not delude themselves about the scale of the challenge, and that robust quantification of what you spend and what you get is much more difficult than anywhere else. Land use change is the second largest source of human greenhouse gas emissions, with deforestation accounting for about 15% of the total. It is also a major driver of biodiversity loss and ecosystems degradation, and has continued at a relentless pace despite scientific warnings that ecosystems such as the Amazon could soon collapse.	
Target Labels: Event Narration, Event Narration, Event Narration, Contextual Details, Event Narration	
Event Narration	The new report emphasizes that protecting tropical forests is crucial for mitigating climate change, preserving biodiversity, and supporting local communities, but it requires a significant increase in funding to achieve these goals.
Event Narration	A significant increase in funding of at least $130 billion annually by the end of the decade is necessary, alongside reductions in beef and dairy consumption and government bans on deforestation, to effectively protect the most at-risk areas of tropical forest.
Event Narration	This funding increase is essential to protect the most at-risk areas of tropical forest, which are currently under threat due to economic incentives driving deforestation.
Contextual Details	Currently, finance to protect forests averages between $2bn and $3bn a year, which is a tiny fraction of the estimated $130bn needed annually.
Event Narration	Accordingly, the report warns that current funding levels are insufficient, and drastic measures are needed to address the issue.

Table 12: Example generation #3

Sentence	Superset Discourse Label
Tina McMahon-Foley is now celebrating her 30th year as a naturalist for Cape Ann Whale Watch (@capeannwhalewatch).',	"Main Event": This sentence directly describes a primary event, noting a milestone in Tina McMahon-Foleyś career as a naturalist, which is the focal point of the document.',
But the story of how she found her way to Cape Ann begins in her former home in Albany, when she worked as a young science teacher in the early 1990s.',	"Previous Event": This sentence describes a specific event that occurred before Tina McMahon-Foley became a naturalist for Cape Ann Whale Watch, providing background information on how she found her way to Cape Ann.',
She was watching television one night when she flipped to the Discovery Channel.',	"Previous Event": This sentence describes a specific event that occurred before Tina McMahon-Foley became a naturalist for Cape Ann Whale Watch, providing context and background information on how she discovered her interest in whales.',
"What came next – a whale documentary, a drive to Massachusetts, a scientist's admiration of her gumption – stick with her today.",	"Consequence": This sentence describes a series of events that directly succeeded a previous event (watching the Discovery Channel) and had a lasting impact on the subjectś life, shaping her current situation as a naturalist.'
On a recent trip, she appeared just as excited to see a whale as she was the first day that scientist, Roger Payne, sent her to sea.',	"Anecdotal Event": This sentence describes a specific, personal experience of Tina McMahon-Foley that illustrates her enduring passion for whale watching, adding an emotional and relatable aspect to her story.',
As a calf breached near the ship, she spoke into the mic to those on board: "Have you caught your breath yet?'	"Anecdotal Event": This sentence describes a specific, personal moment in Tina McMahon-Foleyś experience as a naturalist, which is used to illustrate her enthusiasm and passion for her work, rather than to advance the main narrative of her 30-year career.',

Table 13: An example of the superset of discourse labels that was used to make our discourse schema.