A Novel Multi-Document Retrieval Benchmark Grounded on Journalist Source-Selection in Newswriting

Anonymous ACL submission

Abstract

Multi-document retrieval approaches often overlook the ways different retrievals complement each other when addressing complex queries. In this work, we study journalist source selection in news article writing and examine the *discourse roles* that different sources serve when paired together, finding that narrative function (not simply informational content) is important component of source usage. Then, we introduce a novel IR task to benchmark how well language models can reason about this narrative process. We extract a journalists' initial query and the sources they used from news articles and aim to recover the sources that support this query. We demonstrate that large language models (LLMs) can be employed in multi-step query planning, identifying informational gaps and enhancing retrieval performance, but current approaches to interleave queries fall short. By training auxiliary discourse planners and providing this information to LLMs, we can improve query planning, improving performance on our task by 22% precision and 6.7% F1, all while maintaining recall.

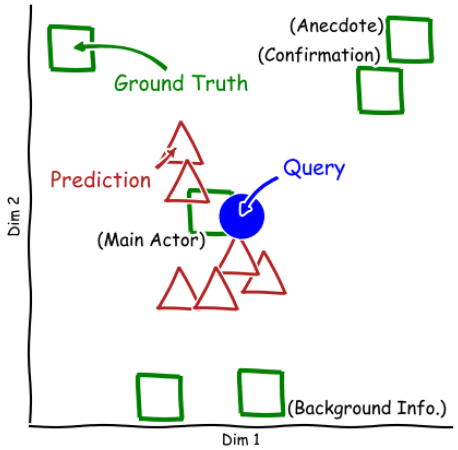

Figure 1: We present a new multi-document retrieval task: source-finding for journalism. We visualize a dataset trial with: a query, ground-truth sources, and vanilla retriever predictions. Notably, we find, ground-truth sources show lower cosine similarity (0.55) to the query than baseline predictions (0.60). We introduce the concept of "retrieval discourse" roles (e.g., "Main Actor" and "Anecdote") highlights how documents complement each other in retrieval.

1 Introduction

Tasks in information retrieval (IR), traditionally focus on retrieving documents based on factual relevance to queries (Manning, 2008), even in approaches that incorporate multi-document retrieval objectives (Zhai et al., 2015; Yu et al., 2023). This overlooks the *discourse* function that different sources of information plays in addressing complex queries (Hearst, 2009). Across a variety of communicative domains – e.g. storytelling (Bruner, 1991), education (ega) and journalism (Tuchman, 1978) – *humans* synthesize information from multiple sources to fulfill different narrative roles. For example, in news articles, it is not enough to cover different subtopics (Zhai et al., 2015); journalists bring together experts, witnesses, and authorities because together, these different sources tell a more complete story (Van Dijk, 1998; Spangher et al., 2024a).

Building off this insight, we introduce a novel IR task that requires us to retrieve multiple information *roles* in answering queries, taking steps towards IR systems that provide more discourse-rich responses (Schank and Abelson, 1977). Specifically, we take a dataset of published news articles (Spangher et al., 2024b) and process them to (1) infer the queries journalists' asked when first starting their writing and (2) extract the sources they used to support their findings, using methods by Spangher et al. (2023). Then, given a large retrieval database consisting of all sources extracted from all articles, our task is to *retrieve the ground-truth set of sources the journalist chose for their article*. Doing this task successfully requires reasoning about the diverse roles and contributions of each

source within a narrative context.

To start, we first seek to better characterize journalistic source selection to gain insights into the discourse behind source selection. We study three kinds of factors driving source-selection: Centrality, Perspective and Narrative Role. While highly central sources provide the bulk of information to news articles, less central sources offer the most varied narrative perspectives. Further, we find that different kinds of stories tend to require different sources.

Next, to model narrative discourse in source retrieval, we utilize LLMs as query planners. LLMs are allowed to issue queries to a retriever and are induced to engage in reasoning threads to issue follow-up queries. As shown in Figure 3, we ask the LLM to reason about the missing pieces of information and formulate queries to retrieve sources providing these diverse informational objectives.

We find that while pretrained LLMs are capable of impressive narrative reasoning, they often tend to miss more interesting sources, mirroring findings by (Spangher et al., 2024b). However, we find that injecting awareness of discourse reasoning typically made by human journalists on similar stories improves retrieval accuracy.

In summary, our contributions are threefold:

- We present a novel IR task grounded in observed sources curated by journalists. This task benchmarks our ability to reason about the different information types that contribute to comprehensive narratives.
- Through extensive analysis, we demonstrate how various sources contribute different elements to a narrative, offering unique viewpoints and fulfilling specific roles within the story's discourse structure. This understanding gives us insights into why certain sources are used together and how they collectively enhance the narrative.
- We show that large language models (LLMs) can be employed in multi-step, interleaved query planning. These models assess initially retrieved sources to identify informational and narrative gaps, guiding subsequent retrieval steps. Incorporating discourse and narrative awareness into the retrieval process significantly improves performance on the task.

Although we focus on news, our focus discourse in retrieval is flexible and taps into core aspects of human narrative development and communication. We seek not only to enhance IR systems' ability to meet complex user needs but also contribute to a deeper understanding of how source-inclusion occurs in narrative structures.

2 Dataset Creation

We utilize a dataset of articles released by Spangher et al. (2024b), which includes 380,000 news articles covering business press releases. The authors originally used this corpus to study the creativity of the sources and the angles in each news article, comparing them to suggestions made by large language models (LLMs). They found that human sources were far more creative than suggestions made by LLMs, indicating that it's a valuable corpus for studying sources.

From this dataset, we sample 50,000 articles and their corresponding press releases that are well-formatted and < 800 words. Our goal is to use the *text* of these news articles to achieve two objectives: (1) create a set of queries and (2) extract a set of informational sources. The queries serve as the input to our retrieval problem, the text of each source-extraction will serve as the ground truth "documents" for each query. All sources are compiled into a retrieval database, which will serve as the database from which we attempt to retrieve the ground truth. Following the definitions in Spangher et al. (2023), sources can be people (e.g., individuals interviewed or issuing statements), documents (e.g., studies, legal documents), or datasets.

Query Generation We provide a large language model (LLM) with both the press release and the corresponding news article, asking it to generate a query that might describe an initial question the journalist had upon reading the press release, which led them to write the article.

Source Creation First, we identify all informational sources in each news article using models trained by Spangher et al. (2023). Then, we then use Llama-3.1-70B[1] to extract, for each source, a complete packet of information pertaining to that source. This includes: describing resolving all coreferences and stating the full names of places, people, and events. This step ensures that we can accurately identify the source later in the retrieval database. Additionally, we extract biographical in-

[1] https://huggingface.co/meta-llama/Meta-Llama-3.1-70B-Instruct/

Query: Is Facebook's (FB) leadership inadequately addressing concerns that moderation policies are applied inconsistently in India, with regards to hate speech from Hindu nationalist politicians?		
Name	Information	Discourse
Ankhi Das (FB Public Policy team)	Opposed internal moves to apply hate-speech rules to a BJP politician and at least three other Hindu nationalist individuals and groups for violating FB's standards.	Main Actor
Former FB employees	A pattern of favoritism exists in India toward the country's ruling party and Hindu hardliners.	Confirmation
FB's Muslim affinity group	Said that Facebook needed to make its policy-enforcement process for high-profile users more transparent and less susceptible to political influence.	Counterpoint
Jeff Horwitz	Wrote an article about FB employees pressing leadership to review its handling of hate speech in India	Background Info.
Newley Purnell	Wrote an article about FB employees pressing leadership to review its handling of hate speech in India.	Confirmation

Table 1: A sample article from our corpus, with query and sources extracted. Labels from our discourse schema (induced from an LLM) is shown in the right column. As can be seen, some sources do not differ greatly from the query (e.g. Former FB employees) while others offer novel dimensions (e.g. Muslim affinity group). Some sources have nearly identical informational content to each other (e.g. Newley Purnell and Jeff Horwitz), and serve to confirm their information.

Discourse Label	**%**	**Discourse Label**	**%**
Main Actor	19.0%	Data	10.2%
Background Info.	18.9%	Confirmation	9.2%
Counterpoint	11.3%	Analysis	7.8%
Anecdotes	10.8%	Broadening	1.6%
Expert	10.5%	Subject	0.7%

Table 2: Distribution of Discourse Types in News Articles. 'Main Actor' and 'Background Info.' are the most common, and 'Subject' the least common.

Centrality	High	Medium	Low
Percentage	21.8%	37.8%	40.0%

Table 3: Percentage of sources by centrality label, queried via LLM.

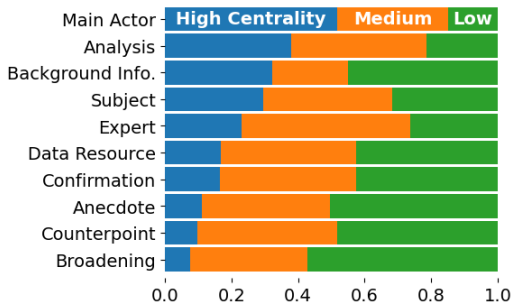

Figure 2: Proportion of sources with each discourse role that occupy a High, Medium or Low Centrality role in their stories.

formation from each source, as well as other metadata, which we will describe in the next section.

2.1 Data Validation

To ensure the validity of our extractions, we take the following steps: We manually annotate a set of 50 documents using a pyramid summarization-like (Nenkova et al., 2007) method: we count the informational units present in each of our extracted sources and then examine the news article to count the units of information attributable to that source. Overall, we find that 87% percent of units in source summaries correspond to units expressed in the original news article. We also manually validate in these 50 documents whether the information in each source stands on its own or if there are unclear coreferences. In 80% of our sources, we are satisfied with the level of detail. Finally, we present two professional journalist a sample of 50 queries and ask them if these queries (1) contain the appropriate level of background information that an experienced journalist would have, and (2) reflect reasonable starting-points for stories. Feedback from journalists confirms that our queries meet these criteria. In total, we extract 400,000 sources, averaging approximately 8.3 sources per document.

3 Analysis

In order to better understand our dataset, we conduct a series of analyses to show how sources are used in news writing by journalists. We express our findings as three primary insights.

Insight #1: Diversity and perspective alone do not characterize source inclusion Maximal coverage (Carbonell and Goldstein, 1998), query diversity (Clarke et al., 2008), and subtopic discov-

ery (Allan, 2003) are common threads in multi-document retrieval: the underlying assumption is that combining diverse sources leads to a more comprehensive retrieval. However, we find that while many sources are chosen for diverse information in news writing, others, are chosen specifically to confirm facts. As shown in Table 1, several sources provide overlapping or identical information (e.g. Jeff Horwitz and Newley Purnell). To investigate further, we compute the pairwise cosine similarity between the SBERT embeddings (Reimers and Gurevych, 2019) of all sources within each article. While the average cosine similarity between source pairs is 0.45 (indicating diversity), a significant minority (7%) of source pairs have a cosine similarity above 0.70, suggesting they provide similar information. This highlights a practice of verification, where multiple sources are used to corroborate facts and enhance credibility (Tuchman, 1972; Kovach and Rosenstiel, 2014).

Traditional models of source selection in journalism often emphasize the inclusion of opposing viewpoints to present a balanced narrative (Gans, 1979). This "for and against" framework assumes that sources are primarily selected to represent different sides of an argument. We conduct an analysis where we score the "perspective" of each source using an LLM [2]. Specifically, we prompt the model to classify the stance of each source as "supporting," "opposing," or "neutral" with respect to the main event or topic of the article (see Appendix B.5 for the full prompt). We find that while some sources do fit into the "for" and "against" categories, this is not universally the case. As shown in Figure 9 in the Appendix, over 30% of sources take an informational perspective without explicitly supporting or opposing any viewpoint. This suggests that source selection is more nuanced than the binary "for and against" model implies. Journalists often include sources to provide context, background information, or expert analysis, which may not directly relate to a polarized viewpoint (Tuchman, 1978).

Insight #2: Sources play clear narrative roles and some narrative roles are more central than others. Given that diversity and stance do not fully explain source selection, we seek alternative ways to characterize how sources are integrated into news stories. We hypothesize that sources play specific narrative roles within the journalistic discourse. To test this hypothesis, we induce a schema to describe the narrative roles played by sources, adapting methods from discourse role learning (Marcu, 2000). We first ask Llama-3.1-70B to generate descriptive labels for the discourse role of each source independently, providing minimal guidance to allow for a broad range of labels. Next, we cluster these labels. We identify eight distinct clusters that represent different narrative roles (e.g., "Main Actor," "Expert" "Background Info."). To validate the reliability of these labels, we recruit two expert journalists to manually annotate a subset of the data. We achieve an 80% agreement rate (Cohen's $\kappa = 0.75$), indicating substantial agreement (Cohen, 1960).

We further investigate how the identified narrative roles correlate with the centrality of sources in news stories. Using Llama-3.1-70B, we label sources based on how central their information is to the main events of the story (see Appendix B.1). Our analysis shows a strong positive correlation ($r = 0.82$) between the centrality of a source and the number of sentences attributed to it, as well as its position in the story (earlier sources tend to be more central). Figure 6 illustrates this.

Insight #3: Certain Kinds of Stories Use Different Kinds of Sources Finally, we examine whether different types of news stories utilize sources differently. We perform clustering analysis on the articles based on the distribution of source narrative roles, using K-means clustering with $k = 8$ clusters determined via the elbow method (Tibshirani et al., 2001).

Our findings indicate that certain clusters correspond to specific genres or topics of news stories. For instance, investigative reports tend to include more "Expert Analysis" and "Background Information" sources, while event coverage focuses on "Main Actors" and "Eyewitnesses." Detailed analysis of these clusters is provided in Appendix B.4, along with examples of stories in each cluster.

This analysis highlights that source selection is context-dependent and varies across different types of journalism. Understanding these patterns can inform the development of more sophisticated information retrieval systems that tailor source recommendations based on the story type.

4 Discourse in Multi-Document Information Retrieval

Given our source and query dataset, described in Section 2, we now present our methodology for

[2] Llama-3.1-70B (Touvron et al., 2023)

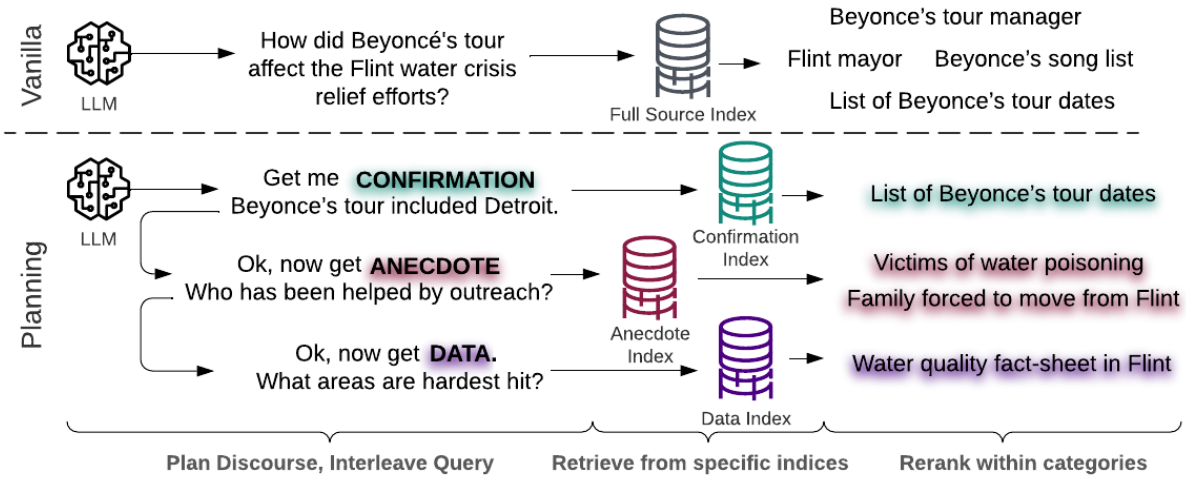

Figure 3: The three-stage discourse-aware retrieval process: (1) Discourse-aware query planning using an LLM with interleaving and discourse role planning, (2) querying discourse-specific indices, and (3) re-ranking retrieved documents within discourse categories.

discourse-aware multi-document retrieval. Motivated by our findings in Section 3, we posit that incorporating discourse structures can significantly enhance the retrieval process. In Section 4.1, we discuss how discourse information can inform the retrieval process and in Section 4.2 we discuss ways to infer a story's discourse requirements.

4.1 Overview of the Retrieval Process

Our retrieval framework consists of three main stages, illustrated in Figure 3: (1) Query Planning, (2) Discourse-Specific Indexing and Retrieval, and (3) Re-ranking. We describe each of these steps, focusing on how discourse roles can be involved.

Stage 1: Query Planning In the first stage, we employ a Large Language Model (LLM) to generate queries $q_1, ...q_n$ sequentially in order to retrieve sources, inspired by Trivedi et al. (2023). Discourse-awareness in this stage means the LLM can reference the discourse role of the source it desires to obtain in query round q_t while generating it's query (we will discuss in Section 4.2 how we infer these discourse roles).

Stage 2: Discourse-Specific Indexing and Retrieval Given a query, q_t, we then retrieve sources $s_1, ...s_k$ relevant to this query. Discourse-awareness in this stage means that the retrieval indices themselves are filtered to discourse roles of sources in our corpus. Traditional multi-document retrieval systems treat all documents equally (Voorhees and Tice, 1999), but our approach organizes the index into hierarchical, discourse-driven sub-indices. This stratification allows for more targeted retrieval. When the LLM generates a query for a particular discourse role, it is directed to the corresponding sub-index.

Stage 3: Re-ranking Finally, given a large set of sources $s_1, ...s_m$ retrieved in the prior steps, we re-rank them to surface the sources that are most relevant together. In this stage, discourse awareness means that we take the most relevant documents *within* each discourse category. This additional layer of categorization prioritizes documents that best fulfill the intended narrative role. We use a re-ranking model that incorporates both relevance and discourse compatibility, similar to the approach in Nogueira and Cho (2019).

4.2 Approaches to Discourse-Aware Planning

As outlined in the previous section, we can incorporate discourse information at each stage in our retrieval process. However, left unexplained was how *we would infer* these discourse roles. Now we discuss the two experimental approaches we take.

Approach #1: Sequential Planning Here, the query-generator is informed of the possible discourse categories, and is asked to pick the next discourse role that a story requires. In other words, at turn t, the LLM views prior $q_{1,...t-1}$ and discourse roles $d_{1,...t-1}$ of retrievals, and is asked to generate the next discourse role, d_t that the story requires.

By allowing an LLM to sequentially generate roles, we hypothesize that we can introduce a human-like planning ability – i.e. often humans

do not know the exact discourse roles a story needs until they get deeper in (Sedorkin, 2015). However, this approach relies the LLM's inherent ability to reason independently about discourse roles without explicit guidance. Prior studies have shown that LLMs struggle with structural reasoning in complex tasks (Spangher et al., 2022), suggesting that this method may be less effective.

Approach #2: A-priori Planning In this approach, we train an auxiliary planner to predict the entire distribution of discourse roles the document will take, a-priori, based on the initial query. To do this, we take the story-clusters discussed in Section 3 (recall, these clusters were generated over discourse distributions), and train a `DistilBERT-base` classifier (Sanh et al., 2019) to *infer* which story cluster a query belongs to.

In other words, the a-prior planner predicts the proportion of each discourse role expected in the final document, based on the initial query. The predicted distribution is then provided to the LLM during the query planning phase[3] We train the auxiliary model on our dataset, achieving a macro F1 score of 0.72 in classifying queries into the correct discourse clusters. The average KL divergence between the predicted and true discourse distributions is 0.7, indicating a close approximation.

Approach #3: Using Ground-Truth Discourse Labels (Oracle) Finally, to differentiate the role of discourse from these two noisy discourse inference techniques, we test an oracle approach. In this approach, we provide the LLM with ground-truth discourse labels extracted during our analysis. By supplying the actual distribution of discourse roles present in the target documents, we assess how well the system can perform when it has perfect knowledge of the sources' discourse structure.

4.3 Experiment Setup

Retriever Model Selection. We choose several representative retrieval methods. These include: (1) *BM25:* a widely-used probabilistic retrieval framework, calculating the relevance of documents to a query based on the frequency of query terms in each document. (2) *Dense Passage Retrieval (DPR)* (Karpukhin et al., 2020): a transformer-based model leveraging dense passage retrieval techniques to effectively capture semantic similarities beyond keyword matching. *DPR* allows us to test whether learned knowledge is more important than planning or reasoning. To finetune DPR, we build a training dataset that including negative samples for in-batch training (Karpukhin et al., 2020). For each positive pair of query q_j and its relevant sources s_j^+, we include n negative tools as negative samples. (3) *SFR*[4]: a 7B text-embedding model developed by Salesforce AI Research that has demonstrated superior performance across multiple benchmarks. We choose SFR as a powerful, large instruction-tuned model in order to understand richer and more nuanced queries that we anticipate our task will require.

LLM Query Agent To implement the LLM agent that is required for planning and reasoning about the discourse needs of stories, we choose Llama-3.1-70B[5] as our LLM for planning. The LLM is provided with discourse planning information with variations described in Section 4.2.

Dataset We perform an 80/20 split for training and test sets. To construct the retrieval index, we aggregate all sources from both sets and organize them according to discourse role, ensuring that each role is indexed separately. For simplicity, we assume that sources maintain the discourse role they had in the original article they were used in. We acknowledge this is a limiting assumption, and perform analysis on sources used multiple times across stories to observe how often the same source has multiple roles. We find that, on average, sources have .43 gini impurity[6], .33 label inconsistency[7], .95 entropy and .55 diversity .55 [8] across discourse roles.

4.4 Results

Our first finding is that, perhaps unsurprisingly, larger retrieval models support better source retrieval. As illustrated in Table 4, the SFR retriever significantly outperforms both BM25 and DPR retrievers across all metrics. Specifically, SFR retriever achieves an F1 score of 26.34% compared

[3]Prompt example: "`We expect this document will contain 50% Background, 30% Expert Analysis, and 20% Main Actor information. Please choose the next discourse role you want to use.`"

[4]https://huggingface.co/Salesforce/SFR-Embedding-2_R

[5]https://huggingface.co/meta-llama/Meta-Llama-3.1-70B-Instruct.

[6]Gini impurity is measured as $1 - \sum_i \left(\frac{l_i}{l_{total}}\right)^2$, where l_i is the count of label i and l_{total} is the sum of all label counts

[7]Inconsistency is defined as $1 - l_{max}/l_{total}$ where l_{max} is the label with the maximum count.

[8]Where diversity is defined as $l_{numunique}/l_{total}$

Retriever	Discourse Strategy		Overall Results			Results by Centrality		
	Sequential	A-priori	Recall	Prec.	F1	High (F1)	Med. (F1)	Low (F1)
BM25	–	–	0.00	0.00	0.00	0.00	0.00	0.00
DPR	–	–	13.98	9.12	11.04	14.42	6.82	5.68
SFR	–	–	**25.81**	27.04	26.34	37.66	22.60	14.37
	✓	–	24.07	25.27	24.60	33.88	21.28	14.05
	–	✓	25.49	31.61	28.04	**40.43**	22.17	14.32
	✓	✓	24.84	**33.15**	**28.12**	40.16	**22.55**	**14.77**
Oracle SFR	–	–	42.77	42.98	42.86	54.02	37.73	26.78

Table 4: We show retrieval strategies and methods in terms of Recall, Precision, F1 score. Each strategy uses multiple retrievers. with the Oracle strategy demonstrating the highest performance metrics.

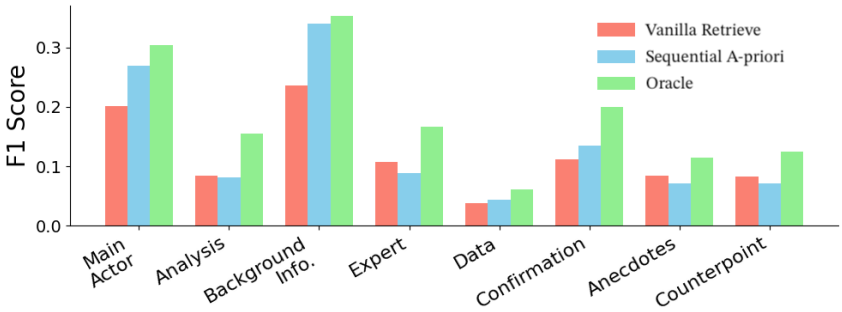

Figure 4: Retrieval accuracy scores, broken down by different discourse types. As can be seen, introducing my discourse planning has a greater impact on certain kinds of discourse categories (e.g. Main Actor and Background Info.) compared with other discourse types (e.g. "Experts", "Anecdotes" and "Counterpoint").

with DPR's F1 score of 11.04% in the same conditions. Substantial improvements underscore the effectiveness of larger embedding models in capturing complex linguistic patterns and semantic similarities, and such advantages that smaller models, even when fine-tuned, fail to match.

Secondly, we observe that an a-prior planning-based approach to reranking *after* querying has a more pronounced impact than sequentially planning. According to the results in Table 4, employing a-priori planning *without* sequential planning[9] yields an F1 score of 28.04%. In contrast, combining both sequential *and* a-prior planning results in a slightly higher F1 score of 28.12%. The small difference between these two trials suggests that a-priori planning alone can substantially enhance retrieval effectiveness, potentially diminishing the incremental benefits introduced by sequential planning. This contrasts with recent results on more conventional conventional QA-based IR tasks, where prompt-based planning strategies were shown to significantly enhance retrieval performances (Trivedi et al., 2022; Huang et al., 2024). We do have to caveat our results with awareness that our a-priori planner was *trained* while our sequential planner relied on LLM pretraining (as did (Trivedi et al., 2022)). This suggests both that (1) a narrative-focused query objective is distinct from purely informational query tasks like those studied previously, and (2) an a-prior plan might contribute positively to these tasks as well, indicating a that there exists templates that journalists follow in choosing sources.

Thirdly, we affirm the role of discourse-level features, such as those captured through planning and clustering, to helping us retrieve sources closer to the ground truth. As evidenced in Table 4, the inclusion of both planning and clustering with the SFR retriever (using 10 candidates) elevates the F1 score from 26.34% to 28.12%. When incorporting oracle discourse information, the F1 score boosts up to 42.86%. This indicates that discourse awareness can provide insights into the structural and relational aspects of the content, which are crucial for understanding context and intent.

5 Discussion

Finally, we investigate why incorporating the discourse aspects into the systems enhances machine's source retrieval ability.

[9] In other words, we simply retrieve $k \times n$-rounds of candidates in the first round, without interleaving, and then re-rank according to the a-priori predicted discourse distribution

Source Centrality As shown in Table 4, the retrieval system shows marked improvement in handling sources of varying centrality when informed by oracle discourse roles, particularly with the oracle setup. For high centrality sources, the Micro-F1 score leaps from 37.66 to 54.02, indicating enhanced effectiveness in identifying and retrieving crucial sources. Similarly, for low centrality sources, the Micro-F1 score rises from 14.37 to 26.78, demonstrating the system's expanded capability to incorporate less central, yet informative perspectives into the narrative, thereby enriching the overall information retrieval process. The improvement from our planning strategies, we observe, originates from the enhanced retrieval of more central sources; this indicates that our planning strategies effectively identifies and prioritizes sources crucial for constructing detailed narratives. However, while the system excels at retrieving high centrality sources, there is room for improvement in capturing more medium and low centrality sources. Enhancing our planning strategies to better include these sources could further enrich the comprehensiveness and diversity of the information retrieval process.

Discourse Role F1 Analysis. As shown in 4, incorporating discourse role information significantly enhances retrieval performance across all source categories, as demonstrated by the results using the oracle discourse role information. By accounting for the specific functions that different sources play in constructing a narrative, the retrieval system becomes more adept at identifying and selecting relevant and comprehensive information. The consistent enhancements across diverse categories highlight the effectiveness of a discourse-aware approach, suggesting that a nuanced understanding of narrative structures is essential for optimizing retrieval outcomes in complex tasks such as multi-document source retrieval.

However, the selective improvements observed with our planning strategies indicate that while these strategies are beneficial, their effectiveness varies across different source categories. Significant gains are achieved in categories central to the narrative—such as Main Actor and Background Information—where the discourse roles are closely aligned with the main query and can be explicitly planned for. This suggests that planning strategies are most effective when the narrative role is straightforward and directly related to the primary focus of the query. In contrast, categories requiring nuanced understanding—such as Analysis, Expert, Anecdotes, and Counterpoint—exhibit less improvement, implying that current planning strategies may not fully capture the complexities inherent in these discourse roles. Consequently, further refinement of these strategies is necessary to enhance retrieval performance in categories that demand deeper contextual and interpretive analysis.

Additional Retrieval Parameters Our preliminary experiments reveal that the effectiveness of discourse-aware retrieval is sensitive to the choice of k, the number of documents retrieved per query. As shown in Figure 5 in the Appendix, the benefits of incorporating discourse information become more pronounced with larger k values. This is consistent with findings from Craswell et al. (2020), who note that re-ranking models have more impact when the initial retrieval set is large. We attempt different methods for learning the ideal k per query: we train a Poisson regression model using a simple Multilayer Perceptron (MLP) on SBERT embeddings (Reimers and Gurevych, 2019). However, the model achieves a low Pearson correlation of $r = 0.35$ between the predicted and actual optimal k values, indicating moderate accuracy. Overall, this additional planning step fails to measurably impact our performance, and we leave further steps in this direction to future work.

6 Conclusion

In this work, we have introduced the concept of discourse in multi-document retrieval tasks, and have framed and introduced a novel task aimed at retrieving sources to assist journalists. We have shown that discourse planning can impact scores, and have introduced two different planners; one based on an LLM and the other based on a learned algorithm. We noted throughout the paper the numerous simplifying assumptions we made in order to implement our task, including: the lack of a trained sequential model, the reliance on ground-truth k and the assumption that sources would retain their initial discourse. We look in future work to more fully return and address these.

7 Limitations

7.1 Ethical Considerations

Our methodology relies on large-scale language models, which have known issues related to bias

and fairness (Sheng et al., 2019; Bender et al., 2021). We take steps to mitigate these concerns by filtering training data for harmful content and evaluating the outputs for biased representations.

7.2 Reproducibility

We provide all code and data necessary to reproduce our experiments at [*GitHub repository link*], following the guidelines set by Pineau et al. (2021) for reproducible research in machine learning. While we provide our code and data in a public repository to promote reproducibility, the computational demands may prevent full replication by those with limited resources. Furthermore, some aspects of our work, particularly the a-priori planning strategy and the LLM's discourse role labeling, involve stochastic elements, which may lead to variations in the results when the models are retrained or fine-tuned on different hardware or datasets.

7.3 Implementation Details

The discourse role classifier and auxiliary planning model are trained with a learning rate of $2e^{-5}$ and batch size of 32.

7.4 Model Limitations

Our study, relying heavily on large language models (LLMs), presents inherent limitations in understanding complex narrative structures. While LLMs such as Llama-3.1-70B are effective at extracting and labeling discourse roles, their performance can be inconsistent when handling nuanced roles like "Anecdote" or "Expert." These roles often require deeper contextual knowledge and interpretative capabilities, which current models struggle to grasp fully. The sequential and a-priori planning strategies we employ only partially mitigate these limitations, leaving room for improvements, particularly in capturing low centrality sources.

7.5 Computational Budget

We conducted our experiments on a combination of BM25, Dense Passage Retrieval (DPR), and SFR-7B embedding models. The SFR model required significant computational resources due to its size (7B parameters). We employed a distributed cluster of 8 NVIDIA A100 GPUs for model training and testing. Fine-tuning the discourse role classifier and auxiliary planner models took approximately 72 hours on this hardware setup. Additionally, large-scale inference, especially with SFR and Llama-3.1-70B, added another 50 hours across multiple processes. This heavy reliance on high-computation hardware restricts the reproducibility of our results for researchers without access to similar resources.

7.6 Data and Annotator Limitations

Our dataset consists of 50,000 news articles sampled from a larger corpus of 380,000, but this sample size may not fully represent the diversity of journalism across various media outlets. Moreover, the annotations for discourse roles were generated using LLMs, and while we manually validated a subset of 50 documents, this represents only a small fraction of the dataset. We involved two professional journalists to assess the validity of our extracted queries and source roles, but this limited human annotation introduces the possibility of bias and errors not being sufficiently captured across the entire dataset.

7.7 Risks and Ethical Considerations

There are several risks associated with the use of LLMs in journalism-related tasks. Firstly, LLMs have known biases, which may inadvertently influence source retrieval, particularly when retrieving contentious or polarized information. Although we filtered the training data to remove harmful content, biases in the models remain a potential issue, especially in politically charged narratives or sensitive topics. Additionally, relying on automated systems for source selection in journalism introduces ethical concerns regarding the transparency of source curation, as these systems may favor certain sources or viewpoints without clear justification.

References

Teaching as story telling: An alternative approach to teaching and curriculum in the elementary school.

James Allan. 2003. Topic detection and tracking: event-based information organization. In *Topic Detection and Tracking*, pages 1–16. Springer.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Jerome Bruner. 1991. The narrative construction of reality. *Critical inquiry*, 18(1):1–21.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings*

of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 335–336. ACM.

Charles LA Clarke, Mallik Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666. ACM.

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

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen Voorhees. 2020. Overview of the trec 2019 deep learning track. In *Proceedings of the Twenty-Eighth Text REtrieval Conference (TREC 2019)*. NIST.

Herbert J Gans. 1979. *Deciding What's News: A Study of CBS Evening News, NBC Nightly News, Newsweek, and Time*. Northwestern University Press.

Marti A Hearst. 2009. *Search user interfaces*. Cambridge university press.

Tenghao Huang, Dongwon Jung, and Muhao Chen. 2024. Planning and editing what you retrieve for enhanced tool learning. *ArXiv*, abs/2404.00450.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Bill Kovach and Tom Rosenstiel. 2014. *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect*, 3rd edition. Three Rivers Press.

Christopher D Manning. 2008. Introduction to information retrieval.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4–es.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. In *arXiv preprint arXiv:1901.04085*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Joelle Pineau, Kyle Vincent, Zaid Barret, and et al. 2021. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Roger C Schank and Robert P Abelson. 1977. Scripts, plans, goals and understanding: An inquiry into human knowledge structures.

Gail Sedorkin. 2015. *Interviewing: A Guide for Journalists and Writers*, 4th edition. Allen & Unwin.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3407–3412. Association for Computational Linguistics.

Alexander Spangher, Matthew DeButts, Nanyun Peng, and Jonathan May. 2024a. Explaining mixtures of sources in news articles. In *Conference on Empirical Methods in Natural Language Processing*.

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

Alexander Spangher, Nanyun Peng, Emilio Ferrara, and Jonathan May. 2023. Identifying informational sources in news articles. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3626–3639.

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

Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

H. Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *ArXiv*, abs/2212.10509.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037.

Gaye Tuchman. 1972. Objectivity as strategic ritual: An examination of newsmen's notions of objectivity. *American Journal of Sociology*, 77(4):660–679.

Gaye Tuchman. 1978. Making news: A study in the construction of reality. *Free Pres*.

Teun A Van Dijk. 1998. *News as discourse*. Routledge.

Ellen M Voorhees and Dawn M Tice. 1999. The trec-8 question answering track report. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 77–82. NIST.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Puxuan Yu, Razieh Rahimi, Zhiqi Huang, and James Allan. 2023. Search result diversification using query aspects as bottlenecks. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3040–3051.

ChengXiang Zhai, William W Cohen, and John Lafferty. 2015. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Acm sigir forum*, volume 49, pages 2–9. ACM New York, NY, USA.

A Appendix

B Analysis of Source Centrality and Perspective in Newswriting

In this section, we explore the role that sources play in newswriting by analyzing two key attributes: *centrality* and *perspective*. Using the Llama-3.1-70B language model, we conducted experiments to label sources based on these attributes and examined how they correlate with the sources' placement and prominence within news articles.

B.1 Centrality of Sources

We employed Llama-3.1-70B to label the centrality of sources in news stories. Centrality refers to how integral a source is to the main narrative of the article. Our hypothesis was that more central sources would not only appear earlier in the articles but also be attributed more sentences.

Figure 6 illustrates the relationship between a source's centrality and its position in the story. The plot indicates that sources labeled as more central tend to appear earlier in the narrative. This suggests that journalists prioritize central sources to establish the main storyline promptly.

Figure 7 shows the number of sentences attributed to sources based on their centrality. The data reveal that more central sources are associated with a higher number of sentences, highlighting their prominence in the discourse of the article.

B.2 Perspective of Sources

In a separate experiment, Llama-3.1-70B was used to label the perspective of sources, categorizing them as either *Informational* or *Authoritative*. We investigated how these perspectives influence the placement of sources in articles and the extent of coverage they receive.

Figure 7 and 7 presents the positions of sources based on their labeled perspective. Both Informational and Authoritative sources tend to appear earlier in the articles. This pattern suggests that sources providing essential information or authoritative insights are introduced promptly to establish context and credibility.

As depicted in Figure 10, Authoritative sources occupy more sentences compared to Informational sources. This indicates that while both types are introduced early, Authoritative sources receive more extensive coverage, possibly due to their perceived expertise and influence on the topic.

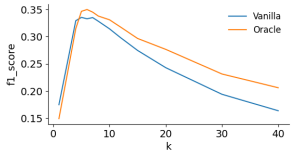

Figure 5: Retrievals benefits of discourse planning grow as k increases relative to baseline.

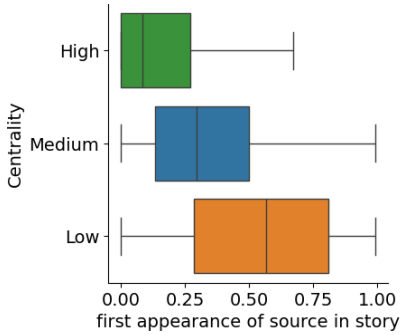

Figure 6: Correlation between centrality assigned to sources by `Llama-3.1-70B` and the first time that source is introduced in the story.

B.3 TopicGPT for Discourse Learning

We train an SBERT-based similarity model, distilled from GPT-4 outputs (OpenAI, 2023), to capture semantic similarities between these labels. We perform hierarchical clustering on the embedding space using the Ward method (Ward Jr, 1963), following the approach of Manning (2008). We then label our full dataset using LLaMA 2 70B, assigning each source to one of the eight categories.

B.4 A-prior Plans: Clustering

As described in the main body, the approach to a-priori planning involved first clustering our label distributions and then training a SequenceClassifier model to predict the cluster, based on the query. We now share more details about the clustering. We clustered KMeans with 8 clusters, cluster centers are shown in Figure 11. Example queries and documents are shown in Tables 6, 7, 8.

B.5 Prompts

Prompt to Score Centrality You will receive a news article and a set of sources to examine in that article.

For each source, provide the following information: (1) Name: who the source is. (2) Perspective: What is their perspective on the main events of the article? Choose as many labels as fit from: ("Authoritative", "Informative", "Supportive", "Skepti-

Label	Definition
Main Actor	Individuals or entities involved in decision-making that effects events in the story.
Subject	Individuals or entities being affected/targeted by events in the story (*i.e.* The converse of "Main Actor").
Anecdote	Real-world stories of people, groups or organizations being affected by events in the story.
Background Info.	Provides broader context to events, helping readers understand the main topic in the context of what is going on and grasp peripheral details.
Broadening	Sources that induce the reader to think about the events of the news article in new or bigger picture.
Analysis	These sources offer insights and forecasts, often explaining what things mean going forward.
Counterpoint	These sources offer diverse perspectives or examples of differences, opposing opinions to provide a more balanced understanding.
Expert	These sources provide essential facts, rules or interpretations to help us understanding the events.
Confirmation	A source whose role is primarily to confirm events that occurred in the news article.
Data Resource	These sources provide statistics and other survey or scientific resources.

Table 5: Definitions for our discourse labeling scheme, generated via LLM-labeling and clustering.

Query: How did news websites handle the expected surge in traffic and demand for video streams during President Barack Obama's inauguration, and what were the consequences for users trying to watch the event online?

Name	Biography	Discourse Label
Daniel Wild	A Web site editor at the New York University School of Medicine	Anecdotes, Examples and Illustration
Akamai	A company that helps many media companies keep up with visitor demand on their Web sites	Data Resource
A Facebook representative	A spokesperson for Facebook	Data Resource
The article's author (unnamed)	The author of the article	Background Information

Table 6: Example of query and ground-truth sources in Cluster #5, the "Data and Resources" cluster.

Query: What are the unintended consequences of receiving a Michelin star, and why would a chef choose to give one up?		
Name	Biography	Discourse Label
Julio Biosca	A chef and owner of Casa Julio, a restaurant in Fontanars dels Alforins, outside of Valencia, Spain, that was awarded a Michelin star in 2009	Main Actor
Julia Perez Lozano	A Spanish food critic	Expert
Frederick Dhooge	A chef and owner of 't Huis van Lede in Belgium	Counterpoint
Skye Gyngell	An Australian chef and owner of Petersham Nurseries Cafe in London	Anecdotes, Examples and Illustration
Gary Pisano	A professor of business administration at Harvard Business School	Analysis
David Munoz	A chef and owner of DiverXo	Anecdotes, Examples and Illustration
The authors of 'Behind the stars: a concise typology of Michelin restaurants in Europe'	Researchers who published a study in the Cornell Hotel & Restaurant Administration Quarterly	Data Resource

Table 7: Example of query and ground-truth sources in Cluster #7, the "Data and Resources" cluster.

Query: What has been the public reaction to Target's decision to allow transgender customers and employees to use the bathroom and fitting rooms that correspond with their gender identity?		
Name	Biography	Discourse Label
Tim Wildmon	President of the American Family Association, a Christian nonprofit organization based in Mississippi	Counterpoint
Kris Hayashi	Executive director at the Transgender Legal Center	Expert
The Williams Institute	An organization that conducted a study on transgender people's experiences with bathrooms	Data Resource
The article's author (unnamed)	The author of the article	Background Information
The American Family Association	A Christian nonprofit organization based in Mississippi	Counterpoint
The Southern Poverty Law Center	An organization that has deemed the American Family Association an extremist group	Counterpoint
Bill Partridge	Oxford Police Chief	Counterpoint
The Human Rights Campaign	An organization that publishes the Corporate Equality Index report	Background Information
Target	The second-largest discount retailer in the nation	Main Actor
The City of Oxford in Alabama	A city in Alabama	Counterpoint
Supporters of the boycott	People who have taken to social media to show their support for the boycott called for in the petition	Counterpoint
Opponents of the petition	People who have shown opposition to the petition	Counterpoint

Table 8: Example of query and ground-truth sources in Cluster #1, the "Counterpoint" cluster.

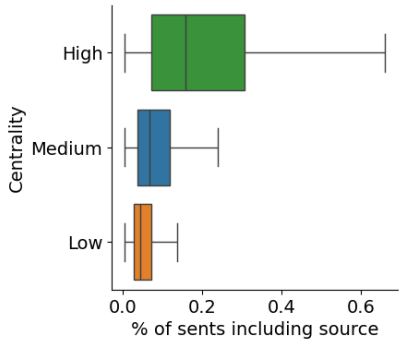

Figure 7: Correlation between centrality assigned to sources by `Llama-3.1-70B` and the percentage of sentences attributed to that source by (Spangher et al., 2023)'s methods.

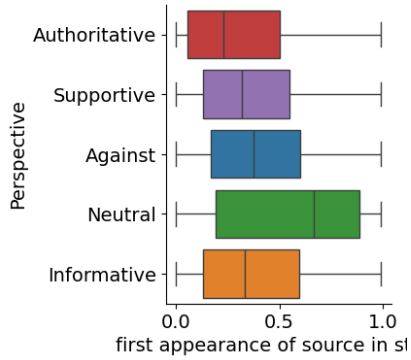

Figure 8: Correlation between perspective assigned to sources by `Llama-3.1-70B` and the first time that source is introduced in the story. Prompts for perspective are shown in Appendix B.5

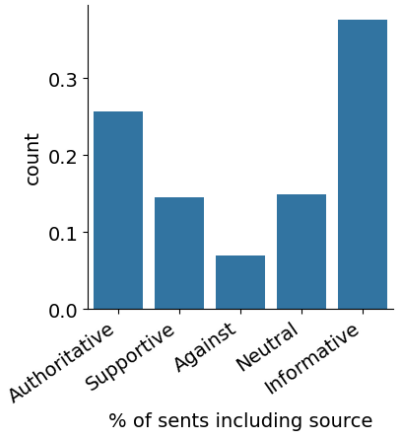

Figure 9: Percentage of sources holding each perspective role, as identified by `Llama-3.1-70B`.

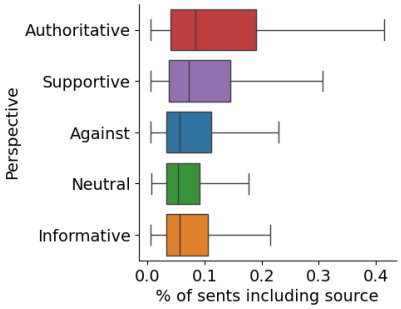

Figure 10: Correlation between perspective assigned to sources by `Llama-3.1-70B` and the percentage of sentences attributed to that source by (Spangher et al., 2023)'s methods.

cal", "Against", "Neutral"). (3) Centrality: How central is this source to the main events of the article? Choose from "High", "Medium", "Low". (4) Is_Error: Did we annotate this source in error? This can happen for many reasons, including if a sentence from the webpage was included in the story unintentionally. Answer with "Yes" or "No".

Here is a news article:

"'{news_article}"'

Please examine the role of each of the following sources:

"'

$$\{target_sources\}$$

"'

For each source, answer the questions above. Output the summary in a list of python dictionaries as in the examples. Don't say anything else.

Prompt to Label Discourse Function You will receive a news article and a set of sources to examine in that article.

For each source in the list, provide the following information, once per source: (1) Name: Exactly copy the name of the source. (2) Narrative Function: Give a generic keyword label to categorize the narrative role the source playes in the article. Infer why the author used the source, and a generalizable statement about the role they play in the article. Don't just summarize their identity. Return in the format: "LABEL": DESCRIPTION.

Here are example outputs. Again, your main task here is to identify a generalizable label that can characterize the narrative role of each source and why the author used them.

[Examples] Example 1:

{{ "Name": "Match Group", "Narrative Function": "Counterpoint: This source is used to compare to the main actor in the news article and provide grounding." }}

Example 2:

{{ "Name": "Dubai Airshow", "Narrative Function": "More Context: This source is used to further

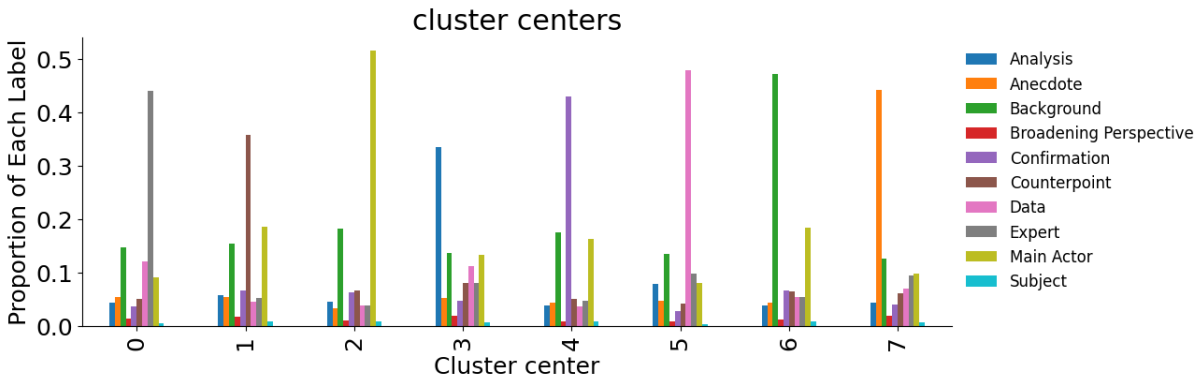

Figure 11: The cluster centers for our KMeans algorithm are distinctive and high-entropy clusters.

expand the context offered and offer a visual setting." }}

Example 3: {{

"Name": "Ann Gough", "Narrative Function": "Victim": This source provides the voice of a user for the product, giving us a personal view of the harm caused by the event. }}

[Instructions]

Now it's your turn. Here is a news article:

“‘{news_article}“‘

Please examine the narrative role of each of the following sources:

“‘

$$\{target_sources\}$$

“‘

For each source, answer the questions above. Output the summary in a list of python dictionaries as in the examples. Don't say anything else.

Prompt to extract source descriptions from news articles You are a helpful news assistant. Here is a news article:
{news_article}

Please summarize each informational source providing information in the article.
Include unnamed or passively expressed sources (e.g. "witnesses", "price signals") if there is information attributable to them.
Include any facts that might have come from the source.
Make sure each source you return refers to just one source. For example: if "John and Jane" both contribute the same information, generate two separate summaries, one for "John" and one for "Jane".
Generate only ONE summary per source.

For each source, provide the following information:
(1) Name: just the name of the source.
(2) Biography: A brief biography of the source mentioned in the article.
(3) Information: Restate the facts provided by the source. Be as SPECIFIC and as VERBOSE as possible.
Contextualize ALL the information the source describes. State the full names of all people, places, events, and ideas mentioned and everything the source says with AS MUCH BACKGROUND INFORMATION from the article so I can fully understand the information the source is giving.
I will look at each source independently without looking at any others, so help me understand the context.

Here are some examples:
example 1:
{ "Name": "Supermarkets around the country",
"Biography": "Retail stores that sell food and other household items",
"Information": "Supermarkets around the country alerted shoppers that prices are likely to continue going up due to the avian flu outbreak, with eggs now average $2.88 per dozen, up 52% since the first confirmed case of avian influenza in February." }

example 2:
{ "Name": "The article's author (unnamed)",
"Biography": "The author of the article",
"Information": "The author stated that Wing, which is collaborating with FedEx and Walgreens on drone delivery, was the first to receive a limited Part 135 certificate. Wing is launching operations in Virginia this month, and the Standard

certification allows UPS to send an unlimited number of drones to the skies, for their cargo load to exceed 55 pounds and for them to fly at night." }

example 3:
{ "Name": "Delta's customers",
"Biography": "People who travel with Delta Air Lines",
"Information": "Delta's customers suggested that they preferred more space on flights amid the COVID-19 pandemic, and they continue to tell Delta that more space provides more peace of mind." }

example 4:
{ "Name": "European Union countries",
"Biography": "Countries that are part of the European Union",
"Information": "European Union countries are working on adopting copyright rules that allow news companies and publishers to negotiate payments with large tech companies like Facebook, Microsoft, and Google that use their content on their platforms." }

Output the summary in a list of python dictionaries as in the examples. Don't say anything else.