A Novel Multi-Document Retrieval Benchmark: 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 discourse function (not simply informational content) is an 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 journalist's 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 incorporating this information into LLMs, we enhance query planning, achieving a *significant* 5% improvement in precision and a 2% increase in F1 score over the previous SOTA, all while maintaining recall.

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 play in addressing complex queries (Hearst, 2009). Across a variety of communicative domains – e.g. storytelling (Bruner, 1991), education (Egan, 1989) 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

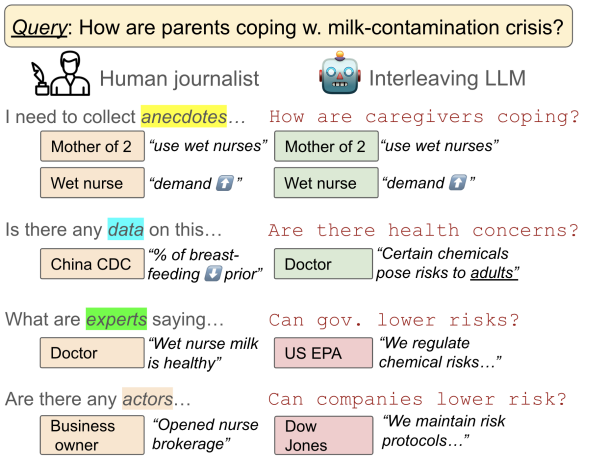

Figure 1: We present a new multi-document retrieval task: source-finding for journalism. Shown above is a complex query, extracted from news articles. On the left are the sources the journalist used to support the query, each annotated with their discourse role. Shown on the right are the queries issued by a baseline LLM interleaved retrieval. Although initially successful (green) the LLM meanders and departs from the original query, retrieving irrelevant sources (red).

(Spangher et al., 2024a). Together, these sources tell a more complete story (Van Dijk, 1998).

Building off this insight, we introduce a novel IR task that requires us to retrieve multiple documents to support complex queries the way a *human* would retrieve them. Specifically, given (1) a dataset of news articles, (2) the initial queries guiding each news article, and (3) sources extracted from all articles, *our task is to retrieve the ground-truth set of sources a journalist chose for the article*. Successfully addressing this requires reasoning about the roles and contributions of each source within a narrative context (Schank and Abelson, 1977).

We start by testing an *interleaving retrieval approach* to (Trivedi et al., 2023) address this task, as shown in Figure 1. In this approach, an LLM is used to iteratively: (1) issue queries to a retriever (2) reason about the sources returned (3)

issue follow-up queries. *However, we find that these interleaved queries quickly repeat, meander, or degenerate, and fail to capture the diversity of sources used in human writing (Section 5).*

We hypothesize that a higher-level planner can guide the interleaving process towards diversity while staying focused on the query. For example, based off the example in Figure 1, we would like a higher-level planner to predict: *this query is likely to answered by "anecdotes", "data", "experts" and "actors"* – we can then use this plan to guide interleaving steps. To make training such a planner tractable, we first constrain the space of possible plans: we do this by developing a novel discourse schema (described in Section 2.2). With this lower-dimensional planning space in hand, we train a high accuracy autoregressive planner.

Finally, we introduce a novel retrieval method called *Planned Interleaved Retrieval (PIR)* to utilize retrieval plans in an interleaving fashion. PIR uses discourse labels in three ways: (1) *querier:* The LLM is given the discourse label for each interleaved query in the prompt. (2) *retriever:* The retrieval database is segmented based on discourse roles. (3) *re-ranker:* The results are reranked within each discourse segment. Taken together, we find that PIR increases retrieval precision by 5% and improves F1 score by 2%.

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 introduce a novel method, *Planned Interleaved Retrieval*, and demonstrate that planning can be used to guide a multi-step, interleaved querying process. Incorporating discourse into the retrieval process, we show, significantly improves performance on the task.

Although we focus on news, our focus discourse in retrieval is flexible, and we have *offered a vision of how retrieval might incorporate higher-level planning structures*. 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 Task and Dataset Creation

To set up our multi-document retrieval task, we wish to create *a large retrieval database where multiple "documents" are labeled as ground-truth for answering each query.* Obtaining gold labels in journalism, though, is challenging: news is experts' domain that is difficult to crowdsource. So, to construct our task, we *reverse-engineer* the *text* of finished news articles, as described below.

2.1 Dataset Creation

For each news article, we extract two items: (1) a query describing the initial question answered by the journalist and (2) the set of informational sources used by the journalist. The queries serve as the input to our retrieval problem, while the text of each source serves as the ground truth matching "document" for each query. 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. We use a dataset of articles released by Spangher et al. (2024b), which includes 380,000 news articles covering business press releases. From this dataset, we sample 50,000 articles and their corresponding press releases. *Press release coverage is a practically useful domain, because press-releases coverage is a necessary and time-sensitive part of business coverage (Petridis et al., 2023).*

Query Generation We provide an 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 Extraction First, we identify all informational sources in each news article using models trained by Spangher et al. (2023). Then, we use Llama-3.1-70B[1] to extract, for each source, a stand-alone packet of information provided by that

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

source[2] "Standalone" means that we can accurately identify the source later in the retrieval database. In total, we extract 400,000 sources, averaging approximately 8.3 sources per document.

2.2 Schema Generation

As described in Section 1, we seek to create a low-dimensional schema to describe our sources (in order to ground our planner). We describe that process now. Inspired by Pham et al. (2024), we first ask an LLM to generate descriptive labels for the discourse role of each source, *based on it's source extraction*. This allows for a broad superset of labels (examples are shown in the Appendix, Table 10.). Then, we cluster these labels by (1) annotating pairs of labels with similarity judgments using an LLM[3], (2) using these annotations to train an SBERT embedding model (Reimers and Gurevych, 2019a), and (3) clustering these embeddings using k-means. We identify eight distinct clusters that represent different narrative roles (e.g., "Main Actor," "Expert" "Background Info"). Definitions for each discourse role are shown in the Appendix, Table 5. Additionally, we ask the LLM to label the centrality of the source: "High" (the source is crucial to the narrative), "Medium" (the source plays a significant role but is not necessary) and "Low" (the source could be easily replaced with another source). We show the breakdown of Discourse Roles by Centrality in Figure 2, and give additional analysis in the Appendix.

2.3 Data Validation

Query and Source Extraction Validation First, we present two professional journalists a sample of 150 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. The journalists confirm 95% of our queries meet these criteria. Next, the journalists manually annotate a set of 396 sources using pyramid summarization evaluation (Nenkova et al., 2007): they count the informational units present in each of extracted source and then examine the news article to count the units of information attributable to that source. Overall, we find that 87% of units in source summaries correspond to units expressed in the original news article. We also manually validate 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.

[2]This includes: describing resolving all coreferences and stating the full names of places, people, and events.

[3]Specifically, whether two different narrative roles generations are substantially the same or not.

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

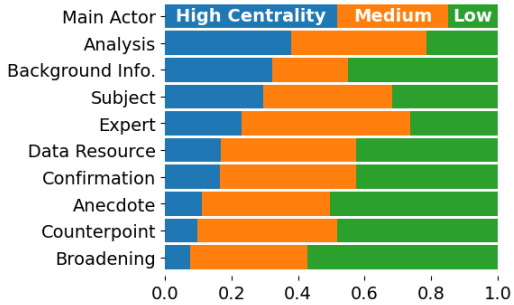

Figure 2: Proportion of sources within each discourse role that occupy High, Medium or Low Centrality in their stories.

Discourse Schema Validation To validate the reliability of these labels, we ask the same two expert journalists to manually annotate the 396 sources with labels from our schema. The journalists achieve a high inter-annotator agreement rate (Cohen's $\kappa = 0.75$) as well as a high agreement rate with our applied labels ($\kappa = 0.64$), indicating substantial agreement (Cohen, 1960).

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: Sources used in multiple documents tend to have the same discourse roles. We expected that sources would often be used in different roles in differet articles: for instance, in Story #1, a police officer might be a "Main Actor", in Story #2 the same police officer might used for "Background info." and in Story #3, for an "Anecdote".

We conduct an analysis on all named sources that

we name-match across two or more articles and find that, on average, sources tend to be classified in the same role (sources have .43 gini impurity[4], .33 label inconsistency[5], .95 entropy and .55 diversity[6] across discourse roles.) This is a crucial insight: for simplicity, in the rest of the paper, we assume that sources' discourse role is only based on their source-text, not the needs of the query.[7]

Insight #2: Diversity and perspective alone do not characterize source inclusion Diversity is a common threads in multi-document retrieval: the underlying assumption is that combining diverse sources leads to a more comprehensive retrieval (Carbonell and Goldstein, 1998; Allan, 2003; Clarke et al., 2008). However, we find that, in news writing, while many sources are chosen for diverse information, others are chosen specifically to confirm facts. For example, ∼10% of sources play a Confirmation role, as in Table 1. We show more analysis in Appendix B.

What other theories exist to explain source-selection criteria in journalism? Gans (1979) suggests that supporting and opposing viewpoints are selected to give a balanced narrative, suggesting that *stance* is a primary driver for source selection. We conduct an analysis of sources' stances in the narrative, using (Ma et al., 2024)[8]. We find that while some sources do fit into the "for" and "against" categories, this is not universally the case. Over 30% of sources take an informational perspective *without explicitly supporting or opposing any viewpoint*[9]. 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).

[4]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

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

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

[7]We acknowledge this is a limiting assumption, and in follow-up work we will remove that assumption. However, because we are already introducing many new ideas and concepts, we hold this constant as it simplifies computation.

[8]Ma et al. (2024) used Llama 3.1 with chain-of-thought prompts to detect stance; this scored highly on popular stance benchmarks. 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 E.4 for the full prompt).

[9]Shown in Figure 9 in the Appendix

Insight #3: Certain Kinds of Stories Use Different Kinds of Sources Finally, we examine whether different types of news stories use sources differently. We manually identify different kinds of coverage: investigative reports, breaking news, etc. (see Appendix E.3 for a full list). We find that different kinds coverage tend to be dominated by different source discourse roles. 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 is provided in Appendix E.3, along with examples of stories. 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 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 *Planned Interleaved Retrieval*

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: Interleaved Querying In the first stage, we employ an LLM to generate queries $q_1, ...q_n$ sequentially in order to retrieve sources, as in 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: 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

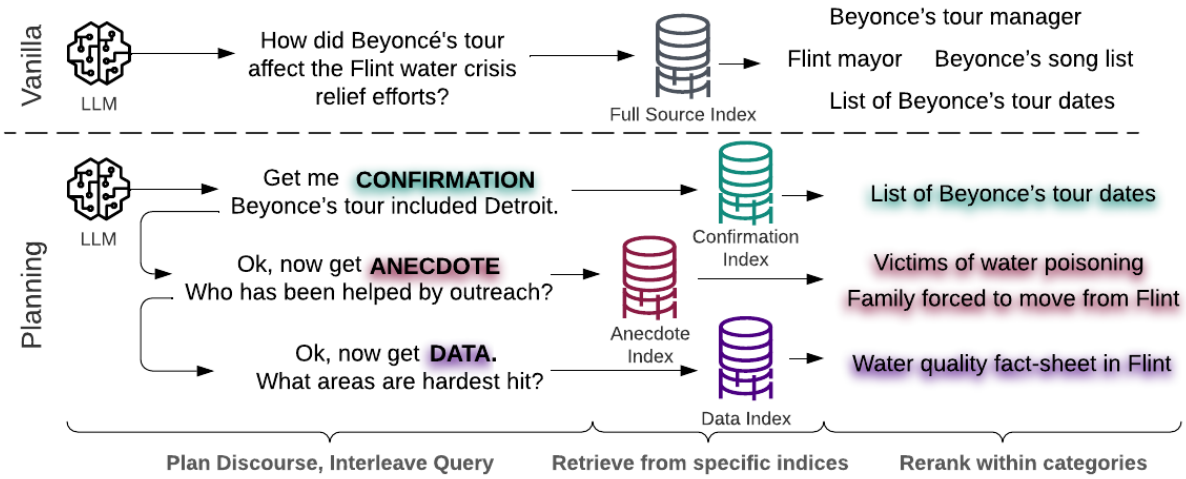

Figure 3: The three-stage discourse-aware retrieval process: (1) Discourse-aware query planning using an LLM with interleaving and discourse role planning, (2) querying discourse-specific indices, and (3) re-ranking retrieved documents within discourse categories.

discourse 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 Two Different Planning Approaches

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 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 cluster articles based on the distribution of source narrative roles, using K-means clustering with $k = 8$ clusters 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[10] 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 be-

[10]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.`"

Retriever	Discourse Strategy		Overall Results			Results by Centrality		
	Sequential	A-priori	Recall	Prec.	F1	High (F1)	Med. (F1)	Low (F1)
BM25 (Robertson and Walker, 1994)			0.00	0.00	0.00	0.00	0.00	0.00
DPR (Karpukhin et al., 2020)			13.98	9.12	11.04	14.42	6.82	5.68
Interleaving (Trivedi et al., 2023)			**25.81**	27.04	26.34	37.66	22.60	14.37
PIR	✓	–	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 PIR	–	–	42.77	42.98	42.86	54.02	37.73	26.78

Table 2: 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. ** indicates significant increases at $p < .01$, obtained via bootstrap resampling ($b = 1,000$).

tween the predicted and true discourse distributions is 0.7, indicating a close approximation.

4.3 Experiment Setup

Retriever. We use SFR[11]: 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 As in Trivedi et al. (2023), an LLM is used to plan and reason about the next query to issue. As in the rest of the paper, we use Llama-3.1-70B.

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, such that each role is indexed separately. *Importantly*: we use one discourse role for each source based on it's text alone, not the query. *We acknowledge again that this is an important limitation,* as noted in Section 3, empirically, most sources maintain similar discourse roles across stories[12]. Maintaining one role simplifies computation for this work: allowing sources to have flexible discourse roles based on the text *and* query is an important future work.

Baselines (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): *we fine-tune* a transformer-based model to to effectively capture semantic similarities beyond keyword matching. Fine-tuned 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) *Interleaving:* we employ SFR with an identical setup to Trivedi et al. (2023) in order to test the ability of LLMs to reason about the needs of the query in the absence of 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.

4.4 Results

Our main finding is that incorporating discourse labels helps us retrieve sources with significantly more accuracy than baseline approaches (we find that these improvements are significant at $p < .01$ by running bootstrapped resamples with $b = 1,000$). As evidenced in Table 2, including discourse labels (with both **a-priori** and **sequential** strategies) elevates the F1 score from 26.34% to 28.12% compared with the baseline *Interleave*. Further, when incorporating oracle discourse information, the F1 score boosts up to 42.86%. This indicates that discourse awareness and planning can provide insights into query needs.

Secondly, and intriguingly, we find that an a-

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

[12] One possible explanation is that journalists observe how other journalists use sources, and use them similarly.

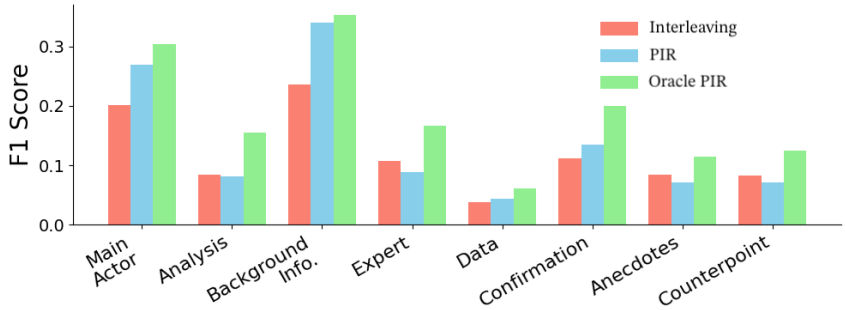

Figure 4: Retrieval accuracy scores, broken down by different discourse types. As can be seen, introducing my discourse planning has a greater impact on certain kinds of discourse categories (e.g. Main Actor and Background Info.) compared with other discourse types (e.g. "Experts", "Anecdotes" and "Counterpoint").

prior planning-based approach has a more pronounced impact than sequential planning. According to the results in Table 2, employing a-priori planning *without* sequential planning[13] 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 QA-based IR tasks, where prompt-based planning strategies were shown to significantly enhance retrieval performances (Trivedi et al., 2023; Huang et al., 2024). These results suggest that our task possesses inherent differences. We do 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., 2023)). 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 is useful in this task, indicating that templates exists that journalists follow.

5 Discussion

We investigate why incorporating the discourse aspects into the systems enhances machine's source retrieval ability above the *Interleaving* approach.

Vanilla Interleaving Tends to Meander : To explain the subpar performance of *Interleaving*, which has shown state-of-the-art results on QA benchmarks, we examine multiple query threads, shown in Appendix A.1. Vanilla interleaving exhibits three notable failure modes. (1) Many queries generated by the planner tend to restate the same objectives or focus on overly narrow aspects of the broader topic without expanding into complementary dimensions (see Appendix A.1, Table 6). This restricts the planner's ability to explore the full range of sources that a humans typically consider (e.g. expert opinions, counterpoints, or data analysis), thus producing a less well-rounded article. (2) Paradoxically, while interleaving often remains closely aligned with the initial query's intent, it also suffers from a tendency to drift when progressing through subsequent queries. For instance, an initial focus on the societal consequences of an issue may eventually lead to highly specific and less generalizable topics that deviate from the core inquiry (e.g. in Figure 1 and Appendix A.1, Table 7). (3) Finally, even when the planner maintains alignment with the initial query, it often fails to explicitly request critical discourse roles, such as expert analyses or contrasting viewpoints (Appendix A.1, Table 8). Consequently, the output of vanilla interleaving lacks the depth and balance.

Source Centrality As shown in Table 2, the retrieval system shows marked improvement in handling sources of varying centrality when informed by 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

[13] 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

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 to better include these sources could further enrich the comprehensiveness of the IR process.

Discourse Role F1 Analysis As shown in Figure 4, incorporating discourse role information significantly enhances retrieval performance across discourse roles. By accounting for the specific functions that sources play in constructing a narrative, the retrieval system is more adept at identifying and selecting *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, 2019b). However, the model achieves a low Pearson correlation of $r = 0.35$ between the predicted and actual optimal k values. Overall, this additional planning step fails to measurably impact performance. We leave further steps to future work.

6 Related Work

Multi-document IR has long recognized that document relevance alone is often insufficient for complex queries (Manning, 2008; Carbonell and Goldstein, 1998; Clarke et al., 2008). Early work centered on keyword-based matching and later evolved into coverage-based paradigms that target subtopic diversity. However, most approaches still optimize solely for factual alignment, overlooking how sources might fulfill different *narrative* or *discourse* functions. Recent work in *LLM-based reasoning* have introduced methods such as *chain-of-thought* (Wei et al., 2022; Trivedi et al., 2023), which encourage models to articulate intermediate inferences improve multi-hop or compositional queries. While *interleaving IR* iteratively refines queries, they largely neglect explicit discourse roles. Our work addresses this gap by designing a discourse-driven selection paradigm, where roles like "expert opinion" or "background info" are explicitly modeled. We show how this lens significantly enriches the set of retrieved documents — an essential step toward tasks that value not just *what* sources provide, but *why* they are chosen.

7 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.

8 Limitations

8.1 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.

8.2 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.

8.3 Implementation Details

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

8.4 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.

8.5 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.

8.6 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.

8.7 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

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

Anonymous. 2023. Palm + rlhf: Training language models to follow instructions. arXiv:2308.xxxxx [cs.CL].

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.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, and et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

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.

Antonia Creswell, Murray Shanahan, and et al. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

Kieran Egan. 1989. *Teaching as story telling: An alternative approach to teaching and curriculum in the elementary school*. University of Chicago Press.

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.

Gautier Izacard, Edouard Grave, and Armand Joulin. 2022. Few-shot learning with retrieval augmented language models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*.

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.

Tushar Khot, Ashish Sabharwal, and et al. 2023. Decomposition-driven reasoning in language models. *arXiv preprint arXiv:2304.xxxxx*.

Takeshi Kojima, Shixiang Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasaki. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Chenfei Liang, Can Wu, et al. 2023. Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis. arXiv:2303.xxxxx [cs.CL].

Junxia Ma, Changjiang Wang, Hanwen Xing, Dongming Zhao, and Yazhou Zhang. 2024. Chain of stance: Stance detection with large language models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 82–94. Springer.

Christopher D Manning. 2008. Introduction to information retrieval.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. OpenAI Blog.

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*.

Laura Parisi, Avinash Athreya, et al. 2022. Talm: Tool-augmented language models. arXiv:2210.xxxxx [cs.CL].

Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–16.

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.

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.

Ofir Press, Libby Barak, and et al. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

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

Nils Reimers and Iryna Gurevych. 2019b. 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.

Stephen E. Robertson and Steve Walker. 1994. Some simple effective approximations to the 2–poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241. Springer-Verlag.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

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.

Timo Schick, Jay Dwivedi-Yu, et al. 2022. Modular reasoning, knowledge and language (mrkl) systems. arXiv:2203.xxxxx [cs.CL].

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.

Adam Passman Shinn, Rami Labash, Daniel Hesslow, and et al. 2023. Reflexion: Language agents with verbal reinforcement learning. arXiv:2303.xxxxx [cs.CL].

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*.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Sydney Ichien, and et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. In *International Conference on Learning Representations (ICLR)*.

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.

Ruiqi Zhang, Linda Li, Xiaodong Liu, Bill Dolan, and et al. 2022. Automatic chain of thought prompting in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Appendix

A.1 Examples of Query Threads

B Further Characterization of Discourse Roles

As shown in Table 3, 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, 2019b) 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).

C Expanded Related Works

D Related Works

In this section, we situate our work at the intersection of *information retrieval (IR)*, *discourse-driven narrative construction*, and *reasoning-based NLP frameworks* (e.g., *chain-of-thought*, *interleaving retrieval*, and *agentic NLP*). We highlight how methods in each of these areas contribute to our proposed task of *multi-document source retrieval for journalism* and illustrate how discourse modeling adds a crucial layer of planning beyond conventional IR objectives.

D.1 Information Retrieval

Information retrieval has a rich history, beginning with classical keyword-based approaches (e.g., TF-IDF, BM25) that treat queries and documents as bags of words (Manning, 2008; Salton and McGill, 1983). These methods remain foundational to modern IR pipelines. Over time, specialized paradigms like *diversified retrieval* (Carbonell and Goldstein, 1998; Clarke et al., 2008) and *subtopic retrieval* (Allan, 2003) have evolved to handle broad, multi-faceted queries by reducing redundancy and maximizing coverage. Our work follows the spirit of *multi-document retrieval* (Zhai et al., 2015), emphasizing that single-document relevance alone is inadequate for tasks requiring multiple complementary sources.

Building upon these foundations, *dense vector retrieval* has emerged, leveraging neural embeddings to map queries and documents into a shared semantic space (Karpukhin et al., 2020; Reimers and Gurevych, 2019a). These approaches excel at capturing deeper lexical and semantic relationships, outperforming bag-of-words techniques in various domains. "Retrieval-augmented" language models further enhance this by prompting LLMs to iteratively refine queries and re-rank candidate documents (Izacard et al., 2022). Our framework extends this line of work by explicitly modeling *discourse roles* rather than purely semantic or topical overlaps, aiming to retrieve sources that complement each other *functionally* in narrative building.

D.2 Planning and Chain-of-Thought Reasoning

Concurrently, large language models (LLMs) have catalyzed progress in *few-shot learning*, *text generation*, and *reasoning*. *Chain-of-Thought (CoT)* prompting (Wei et al., 2022) encourages models to articulate intermediate reasoning steps, improving factual accuracy and multi-hop inference in question-answering (Zhang et al., 2022) and math tasks (Kojima et al., 2022). Our work adapts these insights to *iterative query planning*, where subqueries are tied to distinct *discourse functions*.

Recent research has explored enhanced *planning* frameworks in LLM-driven pipelines. For instance, "self-ask" prompts (Press et al., 2022) or symbolic reasoning modules (Anonymous, 2023) help break down complex tasks. We incorporate and extend these ideas by linking intermediate reasoning steps to specific *discourse roles*—such as "expert perspective" or "main actor"—thereby imposing additional structure on the retrieval process.

D.3 Interleaving Retrieval and Follow-Ups

Interleaving retrieval (Trivedi et al., 2023) describes a process where an LLM iteratively queries a retrieval system, inspects the results, and refines queries for subsequent rounds. Follow-up works build on this paradigm with more advanced *planning modules* (Huang et al., 2024) or specialized *retrieval agents* (Nakano et al., 2021). These techniques aim to systematically explore or fill information gaps across multiple query iterations.

However, existing interleaving methods commonly focus on retrieving the *most relevant* documents. Our work posits that "relevance" alone is insufficient for tasks like *journalistic source selection*, where each source must also fulfill a particular *narrative function*. Accordingly, we propose

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 3: 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.

Centrality	High	Medium	Low
Percentage	21.8%	37.8%	40.0%

Table 4: Percentage of sources by centrality label, queried via LLM.

Planned Interleaved Retrieval, which explicitly encodes discourse roles in a *plan* or distribution of roles needed for a coherent story. This approach reduces the tendency for queries to meander or become repetitive, facilitating *diversity* and *functional complementarity* in the retrieved documents.

D.4 Reasoning in NLP

While NLP systems have historically tackled classification and generation tasks, *multi-step reasoning* is increasingly central to modern challenges (Creswell et al., 2022; Bubeck et al., 2023). Prior work explored neural *memory networks* for logical inference (Weston et al., 2015); LLMs, however, can now articulate more explicit, *symbolic* reasoning steps in few-shot or chain-of-thought paradigms. Yet, even advanced models struggle with tasks requiring strict logical consistency or complex entity tracking across documents (Khot et al., 2023).

Our framework adds a *discourse reasoning* lens to multi-step retrieval. Instead of merely stringing together sub-questions for coverage, we examine *why* different sources are chosen together. We label sources by *narrative role* (e.g., confirming facts, providing an anecdote, serving as an expert) and reason about how each source contributes to the story's completeness. By encoding these discourse intentions, we achieve richer retrieval outcomes aligned with real-world journalistic practices (Tuchman, 1978).

D.5 Agentic NLP

A growing interest in *agentic NLP* frames LLMs as *autonomous agents* that plan, retrieve, and act upon external tools, such as search engines or databases (Shinn et al., 2023; Liang et al., 2023). Architectures like *MRKL* (Schick et al., 2022) and *tool-augmented LLMs* (Parisi et al., 2022) treat the language model as a decision-making orchestrator that delegates subtasks to specialized APIs. Such systems can handle multi-hop QA or web browsing by adaptively issuing queries and integrating results.

Our method can be viewed as a specialized agentic approach, where an LLM "agent" controls a multi-document retrieval pipeline using *discourse-level guidance*. Instead of purely seeking factual coverage, the LLM is tasked with *ensuring* that each source fulfills a *unique narrative function*. By incorporating higher-level organizational structures (i.e., discourse roles) into the agent's plan, we steer retrieval towards more *comprehensive* and *multi-faceted* sets of sources. This approach fits into the broader shift toward agentic NLP, where language models do more than "respond"—they *coordinate* the entire solution process.

Overall, our work is informed by **classical IR** insights on coverage, strengthened by **neural retrieval** methods, and guided by **chain-of-thought**

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.

style planning. We build on the **interleaving retrieval** paradigm but innovate by imposing explicit *discourse structure*, effectively bridging the gap between *unstructured multi-document IR* and *agentic NLP* approaches. By elevating *why* sources are chosen (discourse intentions) alongside *what* they contain (semantic relevance), we deliver more *journalistically valid* and *functionally diverse* retrieval outcomes. The subsequent sections introduce our dataset, experimental setup, and evaluation, illustrating how discourse reasoning substantially improves multi-document retrieval for journalism.

E 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.

E.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.

E.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

Queries
1. What economic and social impacts can the local communities surrounding Mount Everest expect in the aftermath of the earthquake, and how are international aid organizations contributing to the relief and recovery efforts?"
2. What measures is the Nepalese government taking to support the recovery of the climbing industry and affected communities after the earthquake, and how will these efforts impact the long-term sustainability of climbing on Mount Everest?"
3. What specific measures did the Nepalese government take to respond to the earthquake and its impact on Mount Everest, and what lessons have been learned from the disaster to improve future disaster response and preparedness?",
4. What are the long-term plans and strategies being implemented by the Nepalese government and international aid organizations to support the recovery of local communities and the climbing industry affected by the April 2015 Nepal earthquake?"
5. What were the personal experiences and challenges faced by climbers and guides who were on Mount Everest during the April 2015 Nepal earthquake, and how did they respond to the disaster?"
6. What was the impact of the April 2015 Nepal earthquake on Mount Everest climbers, and how would it affect the climbing season?
Retrieved Sources
"A professional mountaineer who was on Makalu, a mountain in the Himalayas southeast of Everest, near the Nepal-China border: Arjun Vajpai said in a video he posted to Facebook from his campsite that there have been a lot of disturbances due to the earthquake, and that there have been a few avalanches in the area, as well as rockfalls, and that most climbers on his team were fine, but he had not yet heard from climbers who moved to other camps."
"The director general of the Nepal Tourism Department: Tulasi Prasad Gautam said that climbing Mount Everest may be canceled this year due to the avalanche, and that the cancellation of the climbing season last year was a key part of Nepal's tourism industry."
"Government officials of Nepal: Nepalese officials said that at least 17 climbers were killed and at least 37 others were injured in the avalanche on Mount Everest."
"A guide at the base camp of Mount Everest and the base camp manager for Asian Trekking: Nima Namgyal Sherpa said the avalanche caused many injuries at the base camp of Mount Everest, and that many camps have been destroyed by the shake and wind from the avalanche, and that all the doctors at the base camp are doing their best to treat and save lives."
"A hiker at the base camp of Mount Everest: Alex Gavan described on Twitter a huge earthquake then huge avalanche that sent him running for life from his tent at the base camp of Mount Everest, and warned that many of the wounded would die if they were not evacuated soon."
"An experienced guide: Ang Sherpa said that about 800 people were already staying at the Everest base camp, which is always a somewhat chaotic collection of tents, equipment and exhausted climbers and their attendants, and that a helicopter rescue operation to the base camp was planned for Sunday morning, when a full tally of the dead and injured should become available."

Table 6: Sample interleaving query for the article www.nytimes.com/2015/04/26/world/asia/everest-climbers-killed-as-nepal-quake-sets-off-avalanche.html, paired with the sources retrieved by this sequence. Many of the sources repeat information, and the queries appear to restate the same objective. Beyond the initial query, subsequent queries primarily pertain to government policies or personal accounts from individuals affected by the disaster. The scope of these queries remains significantly narrower than the range of sources utilized by the human journalist.

Queries
1. What strategies are technology companies, such as Facebook and Twitter, implementing to help users manage their screen time and reduce mindless social media checking, and are these efforts effective?
2. What are the potential long-term health consequences of mindless social media checking, and how have some individuals successfully broken the cycle of excessive social media use?
3. What insights do sociologists or cultural critics offer on the ways in which societal and cultural factors, such as the pressure to present a curated online persona or the fear of missing out, contribute to mindless social media checking, and how can individuals resist these influences to develop healthier social media habits?
4. What strategies or techniques do mental health professionals or behavioral psychologists recommend for individuals to break the cycle of mindless social media checking, and how can these strategies be combined with the features implemented by tech companies to achieve lasting change?
5. What strategies have individuals who have successfully reduced their social media use employed to maintain their new habits over time, and what role do they believe societal and cultural factors play in perpetuating mindless social media checking?
6. How can I, and others, break the cycle of mindless social media checking, which seems to be exacerbated by the constant stream of news and updates from sources like President Trump's Twitter feed?
Retrieved Sources
Experts in the field of neuroscience: Neuroscientists recommend techniques for coping with social media's grip on our brains.
Experts in the field of psychology: Psychologists recommend techniques for coping with social media's grip on our brains.
The President of the United States: Donald Trump's tweets can be annoying and disrupt your workflow, and it's not necessary to see his tweets first thing in the morning.
A technology news website: TechCrunch obtained a Facebook slide deck that shows people check their phones about 100 times a day and Facebook 10-15 times a day.
Surveys and polls conducted by various media outlets: According to a media poll, people don't find news anymore, the news finds them, on Facebook, Twitter, Instagram, and Snapchat.
A newspaper: The New York Post found that the average American checked their phone 80 times per day, or once every 12 minutes, in 2017. One in 10 people surveyed checked it once every four minutes, or about 240 times per day.
People who observe the effects of social media on others: The evidence that we aren't coping well with social media is on display at dinner tables where everyone is staring at screens–and even at crosswalks, where distracted pedestrians walk into traffic.
Social media platforms: Facebook and Instagram will roll out features that show people how much time they've spent on the apps. These features include an activity dashboard that shows how long you're spending on Facebook or Instagram, a daily reminder for when you're hitting your total time for the day, and a way to tune out notifications. These tools can be accessed via the settings page on either app.

Table 7: Sample interleaving query for the article www.wsj.com/articles/take-back-your-brain-from-social-media-1485968678, paired with the sources retrieved by this sequence. This example highlights another failure mode of vanilla interleaving planning: its tendency to drift away from the main query. While Queries 1-5 primarily focus on the consequences of mindless social media usage, Query 6 deviates significantly, introducing highly specific and less generalizable information. This divergence results in poorer retrieval performance.

Queries
1. What role could renewable energy sources, such as solar or wind power, play in replacing the capacity lost due to nuclear power plant closures, and how might policymakers or industry leaders support the development of these alternatives?
2. What are the potential environmental and social consequences of the recent wave of nuclear power plant closures in the US, and how might policy changes or regulatory reforms help mitigate these effects?
3. What are the potential economic and social impacts of the recent wave of nuclear power plant closures on local communities, and how can policymakers or industry leaders mitigate these effects to support workers and local economies?
4. What are the potential national security implications of the recent wave of nuclear power plant closures in the US, and how might the shift towards natural gas and renewables impact the country's energy independence and security?
5. What are the potential health and safety implications of the recent wave of nuclear power plant closures in the US, and how might these closures impact local communities and the long-term stability of the US energy grid?
6. What's behind the recent wave of nuclear power plant closures in the US, and what does this trend mean for the future of nuclear energy in America?
Retrieved Sources
Official documents from the government: The government documents stated that multiple nuclear plants are at risk for early closure and several others have already closed prematurely due to economic circumstances.
A nuclear power plant located in Massachusetts: Pilgrim Nuclear Power Station is closing, with a capacity of 685 megawatts.
A US energy company: FirstEnergy Corp. has said it might decide next year to sell or close its three nuclear plants – Davis-Besse and Perry in Ohio and Beaver Valley in Pennsylvania.
One of the largest providers of nuclear power in the United States: Exelon Corp. announced that it would close two of its nuclear plants in northern Illinois, called Byron and Dresden, in roughly one year's time, even though the plants are licensed to operate for decades more. The plants face revenue shortfalls in the hundreds of millions of dollars. Exelon also said it may need to bring forward the closure dates of two other northern Illinois plants, LaSalle and Braidwood.
A nuclear power plant near Los Angeles: The second and third units of the San Onofre nuclear generating plant near Los Angeles ceased operation in 2013.
A utility company that operates nuclear power plants: Exelon recently announced that it would close the last remaining reactor at the Three Mile Island nuclear plant by 2019 unless policy makers stepped in to support it, due to cheap natural gas cutting regional electricity prices in half and pushing Pennsylvania's nine reactors toward unprofitability.
The industry that generates electricity through nuclear power in the United States: The U.S. nuclear power industry is quietly suffering, with the decline of coal power in the United States making the headlines every week, but the nuclear power industry, which accounts for about 20 percent of U.S. electricity production, is also struggling.

Table 8: Sample interleaving query for the article slate.com/business/2015/10/nuclear-power-is-losing-its-appeal-thanks-fossil-fuels.html, paired with the sources retrieved by this sequence. Vanilla interleaving planning often remains closely aligned with the initial query. In this instance, the queries thoroughly explore the environmental, social, security, and safety implications of nuclear power plant closures. However, they fail to explicitly request expert opinions, data, or counterpoints, critical source types essential for constructing a comprehensive news article.

Queries
1. What are the marketing and business strategies behind EVA Air's decision to introduce the Hello Kitty jet on the San Francisco-Taipei route, and how does this fit into the airline's overall brand and growth plans?
2. What do passengers who have flown on EVA Air's Hello Kitty jets think of the experience, and how does it compare to other themed flights or regular flights?
3. What are the key demographic groups that EVA Air is targeting with its Hello Kitty jet on the San Francisco-Taipei route, and how does the airline plan to tailor its marketing efforts to effectively reach and engage with these groups?
4. What demographic groups are EVA Air targeting with its Hello Kitty jet on the San Francisco-Taipei route, and how does the airline expect this unique offering to differentiate itself from competitors and attract new passengers?
5. What specific marketing and promotional efforts is EVA Air planning to undertake to attract its target demographic groups to the Hello Kitty jet on the San Francisco-Taipei route, and how does the airline expect this unique offering to impact its brand awareness and passenger loyalty in the North American market?
6. What's behind EVA Air's decision to bring its popular Hello Kitty jet to the San Francisco-Taipei route, and what can passengers expect from this unique flying experience?
Retrieved Sources
The founder of TravelSkills.com: The author stated that EVA Air made a media splash several years ago when it decided to cover its jets with cute animated characters created by Japan's Sanrio Company, the most popular of which is Hello Kitty, and that the Hello Kitty planes bring the theme to more than just their paint jobs, with passengers getting specially designed boarding passes and baggage stickers at check-in, and seeing Hello Kitty figures on the flight attendants' aprons, the headrest covers, pillows, tissues, hand cream, napkins, paper cups, and even utensils and snack packaging.
A major Taiwanese airline: For the rest of April, Taiwan's EVA Air will operate three weekly Taipei-SFO flights, three to Seattle and three or four a week to LAX, all with 777-300ERs.
Taiwan's flag carrier: China Airlines plans to increase capacity between San Francisco International and Taipei on May 2, boosting its schedule from seven flights a week to nine, using 777-300ERs.
A Taiwanese airline: EVA Air provided delightful flights and service to the author, unlike their experience on Asiana Airlines.
A Singaporean airline: Singapore Airlines plans to pull its extra-long-range A350-900ULRs off the San Francisco route next fall. The ULR version – which has only business class and premium economy seating – currently operates three of the airline's 10 weekly SFO-Singapore non-stop flights; the others use a standard three-class A350-900, and that model will be used for all 10 of the airline's weekly SFO non-stop flights beginning in October of next year.
A Chinese airline: Air China is planning to boost frequencies on its San Francisco-Beijing route effective March 31 from seven flights a week to 10. The three extra flights will be operated with a 787-9 Dreamliner, while the regular daily service will use a 747-8.

Table 9: Sample interleaving query for the article www.sfgate.com/travel/article/Hello-Kitty-jet-san-francisco-13788445.php, paired with the sources retrieved by this sequence. All of the queries primarily focus on the effects of the new marketing campaign on the airline's passengers, neglecting other important information needs such as data, analysis, or background context. This highlights the vanilla planner's lack of creativity and strategic planning capabilities.

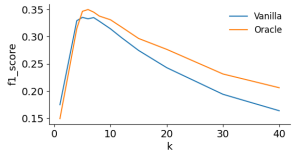

Figure 5: Retrievals benefits of discourse planning grow as k increases relative to baseline.

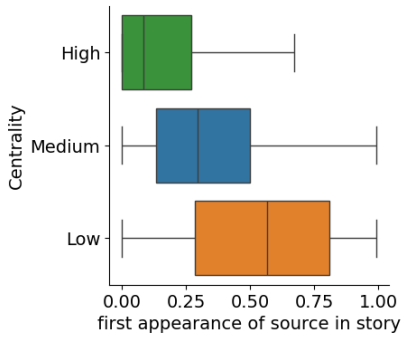

Figure 6: Correlation between centrality assigned to sources by `Llama-3.1-70B` and the first time that source is introduced in the story.

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.

E.3 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 11, 12, 13.

E.4 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", "Skeptical", "Against", "Neutral"). (3) Centrality: How central is this source to the main events of the article? Choose from "High", "Medium", "Low". (4)

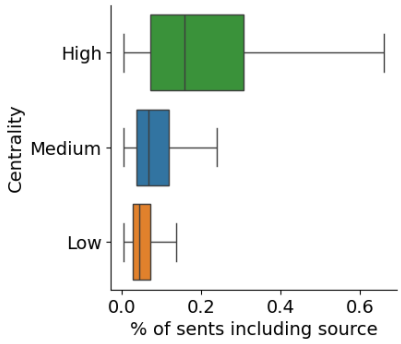

Figure 7: Correlation between centrality assigned to sources by `Llama-3.1-70B` and the percentage of sentences attributed to that source by (Spangher et al., 2023)'s methods.

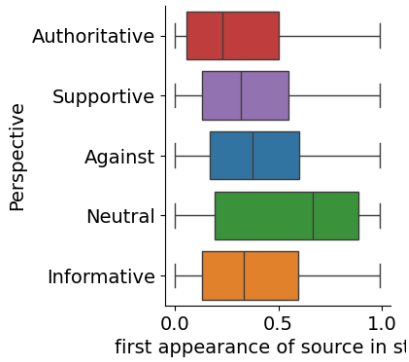

Figure 8: Correlation between perspective assigned to sources by `Llama-3.1-70B` and the first time that source is introduced in the story. Prompts for perspective are shown in Appendix E.4

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

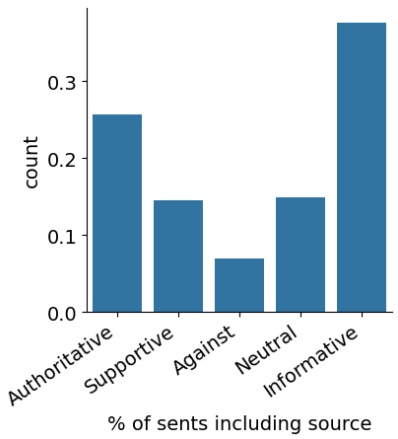

Figure 9: Percentage of sources holding each perspective role, as identified by `Llama-3.1-70B`.

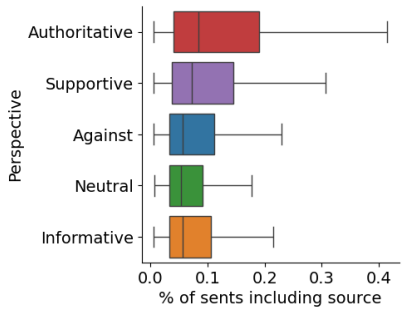

Figure 10: Correlation between perspective assigned to sources by `Llama-3.1-70B` and the percentage of sentences attributed to that source by (Spangher et al., 2023)'s methods.

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 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

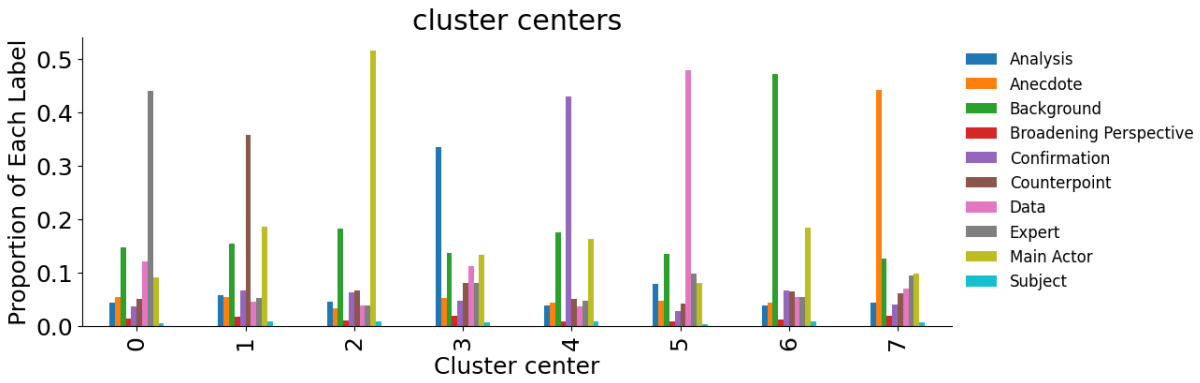

Figure 11: The cluster centers for our KMeans algorithm are distinctive and high-entropy clusters.

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.

	Source Text (to embed)	Narrative Function	Discourse
The FBI	The Federal Bureau of Investigation: The FBI shows that 82 percent of white homicide victims were killed by other white people and 15 percent of white homicide victims were killed by black people	"Fact Checker": This source can provide accurate information and debunk the false statistics.	Data Resource
The U.S. Securities and Exchange Commission	A regulatory agency responsible for enforcing federal securities laws and regulating the securities industry: The U.S. Securities and Exchange Commission has postponed a decision on whether to allow the listing of an exchange-traded fund backed by Bitcoin...	"Authority": This source can be used to establish the regulatory framework and provide the official decision."	Main Actor
The Privacy Rights Clearinghouse	An organization that provides information on how to deal with security breaches: If you learn of a breach involving your driver's license information, contact the agency (in this case the state Department of Driver's Services)	"Authority": This source can be used to provide expert advice and recommendations on how to deal with the data breach.	Expert
CNN	A news organization: Both the Pfizer/BioNTech and Moderna vaccines use an mRNA platform and are well tolerated and safe. Moderna was estimated to be 36.8% effective against symptomatic disease for kids 2-to-5 years of age...	"Providing Data": This source can be used to provide data and statistics to support the claims made in the article.	Data Resource
Tech Crunch	A technology news website: According to a report by Tech Crunch, Detroit Mayor Mike Duggan said on Wolf Blitzer's show on CNN that the city of Detroit received the test kits manufactured by Abbott on April 1.	"Secondary Source": This source can be used to provide additional information and context to the main story.	Background Information
Anil Agarwal	The chairman of the Vedanta group: Anil Agarwal recently said that the group is scouting for more energy and metal assets across India, which includes coal, oil and iron ore.	"Company Strategy": This source can be used to provide insight into the strategy and interests of the Vedanta group.	Analysis
Experts	Unspecified experts in the field of economics or education: Some experts have raised concerns that forgiving student loans may effectively penalize people who already paid off their debt, often while making considerable financial sacrifices.	"Counterpoint": This source can be used to raise concerns and questions about the fairness and effectiveness of the policy."	Counterpoint

Table 10: An example of sources randomly selected from our retrieval database. We show the narrative function originally labeled by Llama-3.1 along with the discourse label applied after clustering. Note how initial narrative function label applied by the LLM narrative function doesn't always align with the final label: for example, row #2 and #3 are both labeled "Authority", however #2 is a more active participant while #3 is more an expert.

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 11: 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
Authors of 'Behind the stars...'	Researchers who published a study in the Cornell Hotel & Restaurant Administration Quarterly	Data Resource

Table 12: 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
Williams Institute	An organization that conducted a study on transgender people's experiences with bathrooms	Data Resource
Article's author	The author of the article	Background Information
American Family Association	A Christian nonprofit organization based in Mississippi	Counterpoint
Southern Poverty Law Center	An organization that has deemed the American Family Association an extremist group	Counterpoint
Bill Partridge	Oxford Police Chief	Counterpoint
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
City of Oxford	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 13: Example of query and ground-truth sources in Cluster #1, the "Counterpoint" cluster.