Uncovering the Bigger Picture: Comprehensive Event Understanding Via Diverse News Retrieval

Anonymous ACL submission

Abstract

Access to diverse perspectives is essential for understanding real-world events, yet most news retrieval systems prioritize textual relevance, leading to redundant results and limited viewpoint exposure. We propose **NEWSCOPE**, a two-stage framework for diverse news retrieval that enhances event coverage by explicitly modeling semantic variation at the sentence level. The first stage retrieves topically relevant content using dense retrieval, while the second stage applies sentence-level clustering and diversity-aware re-ranking to surface complementary information. To evaluate retrieval diversity, we introduce three interpretable metrics, namely *Average Pairwise Distance*, *Positive Cluster Coverage*, and *Information Density Ratio*, and construct two paragraph-level benchmarks: **LocalNews** and **DSGlobal**. Experiments show that **NEWSCOPE** consistently outperforms strong baselines, achieving significantly higher diversity without compromising relevance. Our results demonstrate the effectiveness of fine-grained, interpretable modeling in mitigating redundancy and promoting comprehensive event understanding. *We will release both the benchmark and code upon publication.*

1 Introduction

The ability to access diverse perspectives is fundamental for understanding complex real-world events. However, most existing systems prioritize textual relevance, often producing results that are not only biased toward the query but also highly similar to one another (Kulshrestha et al., 2019). This combination of bias and redundancy limits users' exposure to alternative viewpoints (Spinde et al., 2022). As algorithm-driven news consumption becomes mainstream, relevance-focused retrieval risks amplifying information cocoons and echo chambers. This raises a critical question: *how can retrieval systems surface distinct, complementary perspectives rather than reiterate the same narrative across sources?*
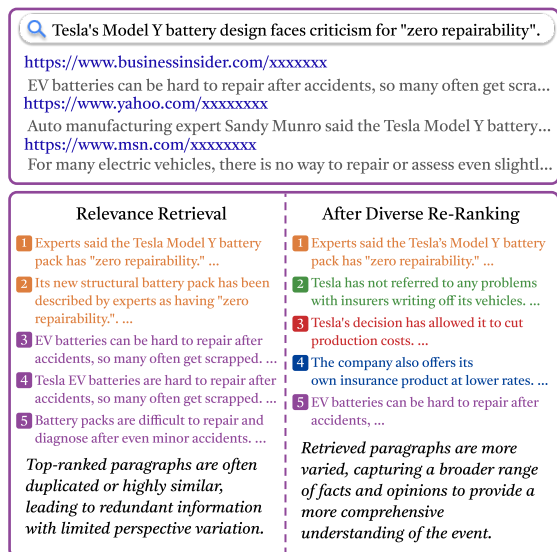

Figure 1: Comparison of dense and diverse retrieval results. Dense retrieval returns similar and redundant paragraphs, while diverse retrieval captures relevant yet distinct paragraphs, improving coverage of different perspectives. Same color indicates semantic similarity.

Traditional text retrieval models, including BM25 (Robertson et al., 1995) and dense retrievers (Karpukhin et al., 2020; Thakur et al., 2021), prioritize textual relevance, often favoring lexically similar documents and repeating prominent entities (Fayyaz et al., 2025). While effective for topical matching, they yield semantically overlapping results. Re-ranking methods like MMR (Carbonell and Goldstein, 1998) introduce diversity penalties, but operate at a coarse, article-level embedding representation and fail to capture fine-grained differences in perspectives, framing, and coverage. As shown in Figure 1, relevance retrieval methods often produce near-duplicate content, limiting comprehensive understanding of events from complementary viewpoints.

Diversified retrieval has been explored in recom-

mendation systems (Yu et al., 2014; Chen et al., 2018; Zhang et al., 2023), where diversity is defined via user behavior or product categories. However, such domain-specific notions do not directly transfer to news retrieval, where diversity involves distinct factual aspects and viewpoints. Existing methods (Xia et al., 2016; Jiang et al., 2017) are typically item-level and supervised, limiting interpretability and scalability. These limitations give rise to two core research questions:

RQ1: *How can we systematically define and measure diversity in news retrieval?*

RQ2: *How can we balance high relevance with comprehensive coverage of distinct perspectives?*

To this end, we introduce the task of **event-centric diverse news retrieval**, which aims to retrieve relevant documents that collectively represent multiple comprehensive perspectives on a target event. We treat sentences as the atomic units of semantic meaning and model diversity at the sentence level to capture fine-grained distinctions. We operate retrieval at the paragraph level to ensure output coherence. To evaluate retrieval diversity, we propose three interpretable metrics: *Average Pairwise Distance* (D), *Positive Cluster Coverage* (C), and *Information Density Ratio* (I), which capture perspective coverage and semantic richness beyond standard precision and recall.

Motivated by these challenges, we propose **NEWSCOPE** (*NEws Understanding via Weighted Sentence Clustering and Optimization for Perspective Enrichment*), a two-stage diversified retrieval framework. The first stage efficiently retrieves candidate documents based on relevance scoring. The second stage applies sentence-level semantic clustering and diversity-aware re-ranking to construct a set of paragraphs that are both relevant and perspectively diverse. To support the task, we construct two paragraph-annotated benchmarks: **LocalNews**, focused on regional reporting, and **DS-Global**, an adaptation of the DiverseSumm summarization corpus (Huang et al., 2024a) focused on global news for retrieval evaluation. Experiments show that **NEWSCOPE** consistently outperforms strong baselines, achieving significantly higher diversity without compromising relevance, while maintaining practical efficiency.

Our main contributions are as follows:

- We formalize the task of **event-centric diverse news retrieval** and propose three novel

interpretable metrics to evaluate the diversity of retrieved content, capturing fine-grained information coverage.

- We propose **NEWSCOPE**, a two-stage retrieval framework that enhance dense relevance retrieval with sentence-level clustering and diversity-aware re-ranking to promote comprehensive perspective coverage.

- We curate two paragraph-level benchmarks, **LocalNews** and **DSGlobal** to support evaluation. We show that NEWSCOPE consistently outperforms strong baselines on both datasets.

Our work highlights the importance of fine-grained and interpretable diversity modeling for news retrieval, aiming to break echo chambers and support more comprehensive, unbiased event understanding. While we focus on the news domain where divergent perspectives and conflicting narratives are common, the proposed framework is domain-agnostic and can be extended to other contexts where diverse information is valuable.

2 Related Work

2.1 Comprehensive News Coverage

Comprehensive news understanding requires aggregating perspectives from multiple sources. Traditional fact-checking and misinformation detection work has focused on claim-level veracity (Thorne et al., 2018), often ignoring source bias. Recent studies warn of selective exposure to biased narratives (Estornell et al., 2020; Fung et al., 2022; Singamsetty et al., 2023; Rodrigo-Ginés et al., 2024; Wang et al., 2023). Retrieval-augmented generation (RAG) pipelines have been adopted for evidence-backed verification (Hu et al., 2023; Singhal et al., 2024; Sriram et al., 2024; Russo et al., 2024), yet these focus on factual correctness over viewpoint diversity.

Recent work on social media analysis has recognized that capturing a diverse range of viewpoints is essential for understanding the full spectrum of narratives (Sundriyal et al., 2024; Agmon et al., 2024; Chuai et al., 2024). Some studies focus on aligning sentence-level information across different perspectives or generating summaries of news events but do not incorporate retrieval mechanisms. The methods often assume access to balanced inputs for each event and lack explicit retrieval components. (Jaradat et al., 2024; Chen et al., 2024;
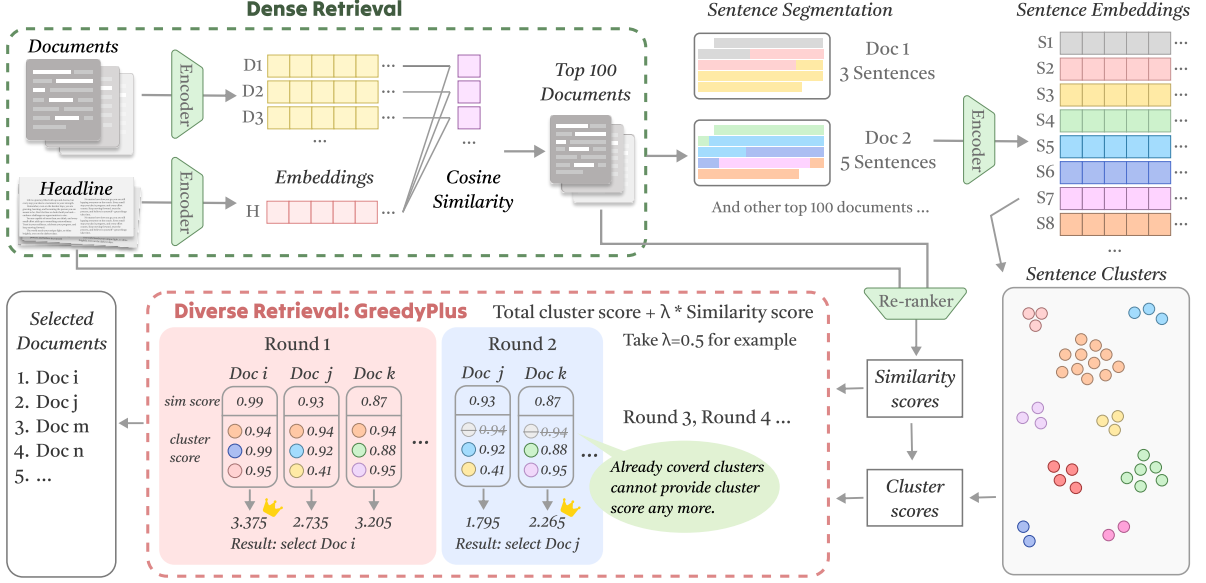
2

Figure 2: Overview of the **NEWSCOPE** framework. The two-stage pipeline combines efficient dense retrieval with sentence-level diversity-aware re-ranking to ensure both relevance and comprehensive perspective coverage.

Zhang et al., 2024; Huang et al., 2024a; Fabbri et al., 2019), which motivate the need for retrieval strategies that ensure balanced and diverse event representations.

2.2 Diverse Retrieval

Traditional retrieval models such as BM25 (Robertson et al., 1995) and dense retrievers (Thakur et al., 2021; Muennighoff et al., 2023) prioritize textual similarity, often surfacing redundant content that repeats dominant narratives. Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) introduces a relevance–diversity trade-off and has been adapted for summarization and dialogue tracking (Fabbri et al., 2019; King and Flanigan, 2023), but it lacks explicit modeling of fine-grained semantic coverage and interpretability.

Diversity-oriented retrieval has been widely explored in other domains(Yu et al., 2014). DPP-based models (Chen et al., 2018) and disentangled representations (Zhang et al., 2023) are used in recommendation systems to reduce redundancy across items, but they rely on user profiles such as purchase history and are evaluated on e-commerce, music, or video platforms, making them less applicable to news retrieval. Other approaches (Xia et al., 2016; Jiang et al., 2017) rely on supervised models and subtopic attention mechanisms, requiring large amount of labeled data for training.

Within news retrieval, DkMIPS (Huang et al., 2024b), as used in DiversiNews (Sun et al., 2024), promotes stance diversity by penalizing similarity

in dense embeddings. It struggles with fine-grained diversity control and lacks semantic transparency. PerspectroScope (Chen et al., 2019) targets claim-centric stance discovery by identifying supporting and opposing views. In contrast, our setting is event-centric and stance-agnostic.

We introduce an unsupervised, interpretable framework for news retrieval. It jointly optimizes relevance and sentence-level diversity without requiring domain-specific training or annotated stances, making it more generalizable across real-world news contexts.

3 Problem Formulation

Given a target event E and a large news corpus $\mathcal{C}$, the goal of **event-centric diverse news retrieval** is to retrieve a set of k paragraphs $\mathcal{R} = \{r_1, r_2, \ldots, r_k\}$ from $\mathcal{C}$ such that:

Relevance: Each paragraph $r_i \in \mathcal{R}$ is topically relevant to the event E.

Diversity: The set $\mathcal{R}$ collectively covers semantically distinct aspects or perspectives of E, minimizing redundancy and maximizing subtopic coverage.

We treat paragraphs as the retrieval unit and model diversity at the sentence level to capture fine-grained variation between results. This enables the retrieval of complementary viewpoints that enhance the comprehensiveness of information presented to the user.

3

4 The NEWSCOPE Framework

We propose **NEWSCOPE**, a two-stage framework designed to retrieve relevant news content that collectively reflects a diverse set of perspectives on a given event. The first stage performs relevance-based retrieval using dense embeddings to collect candidate paragraphs. The second stage performs diversity-aware reranking by modeling sentence-level semantic variation and selecting content that maximizes both coverage and informativeness. Figure 2 provides an overview of the framework.

4.1 Stage I: Relevance-Based Retrieval

Given an event headline, we first retrieve a candidate set of paragraphs using a dense embedding-based retriever. Specifically, both the headline and candidate paragraphs are encoded using the `bilingual-embedding-large`[1] (Thakur et al., 2020) model. Paragraphs are ranked by their cosine similarity to the headline in the embedding space. This step ensures that only highly relevant content is forwarded to the reranking stage, filtering out unrelated or low-quality paragraphs.

4.2 Stage II: Diversity-Aware Reranking

To capture diverse viewpoints, we rerank the candidate paragraphs by explicitly modeling sentence-level semantic diversity. This stage comprises three steps: (1) sentence clustering to identify semantically similar content, (2) greedy selection to cover distinct clusters, and (3) optional cluster weighting to jointly optimize diversity and relevance.

4.2.1 Sentence Clustering

We segment each paragraph in the candidate set into individual sentences, treating sentences as atomic meaning units. Each sentence is encoded as a dense vector. To uncover semantic similarity, we apply the OPTICS clustering algorithm (Ankerst et al., 1999), which groups semantically related sentences based on density without requiring a predefined number of clusters. This yields a set of sentence clusters that serve as proxies for distinct informational aspects of the event.

4.2.2 Greedy Cluster Selection (GreedySCS)

To promote diversity, we employ a greedy selection strategy that iteratively selects paragraphs covering the most novel sentence clusters. Let P be the set of candidate paragraphs and C the set of sentence

---

Algorithm 1 Greedy Cluster Selection

Require: P: Set of candidate paragraphs; C: Sentence clusters
1: Initialize $S \leftarrow \emptyset, U \leftarrow C$
2: **while** $|S| > k$ **or** $|U| >$ coverage_threshold **do**
3:    **for** $p \in P$ **do**
4:       Score$(p) \leftarrow |U \cap \text{Clusters}(p)|$
5:    **end for**
6:    $p^* \leftarrow \arg\max_p \text{Score}(p)$
7:    $S \leftarrow S \cup \{p^*\}, U \leftarrow U \setminus \text{Clusters}(p^*)$
8: **end while**
9: **return** S

---

clusters. We initialize the selected paragraph set $S \leftarrow \emptyset$ and the uncovered cluster set $U \leftarrow C$. At each step, we score each paragraph by the number of uncovered clusters it contains and greedily select the one with the highest score. Formally,

$$\text{Score}(p) = |U \cap \text{Clusters}(p)|,$$

where Clusters(p) denotes the cluster indices that the contains at least one sentence from the paragraph, i.e. $\text{Clusters}(p) = \{C_{s_i} \mid s_i \in p\}$. The selected paragraph is added to S, and its corresponding clusters are marked as covered. The process continues until a stopping condition is met: either a predefined number of paragraphs is selected or sufficient cluster coverage is achieved. This process is outlined in Algorithm 1.

This method, referred to as **GreedySCS**, mitigates information redundancy and promotes a broad representation of perspectives by prioritizing novel clusters.

4.2.3 Cluster-Based Weighting (GreedyPlus)

To further refine paragraph selection, we enhance GreedySCS with a soft weighting mechanism that jointly models relevance and diversity. Specifically, we assign an importance score to each sentence cluster based on its overall relevance to the query.

Cluster Score For each sentence cluster c, we compute a relevance-based weight by averaging the similarity between the headline h and all paragraphs containing sentences from that cluster:

$$\text{ClusterScore}(c) = \frac{1}{|c|} \sum_{s_i \in c} \text{Sim}(h, p_{s_i}),$$

where s_i is a sentence in cluster c, p_{s_i} is the paragraph containing s_i, and $\text{Sim}(h, p_{s_i})$ is the similarity score computed using the `bge-reranker-large` model[2]. $|c|$ is the number of sentences contained in this cluster.

---

[1] https://huggingface.co/Lajavaness/bilingual-embedding-large

[2] https://huggingface.co/BAAI/bge-reranker-large

The cluster score dynamically weights the importance and relevance of sentence clusters with respect to the query headline, serving as a fine-grained, soft noise filter that downweights low-quality or off-topic clusters present in the candidate pool from Stage I, such as incidental content or promotional text.

Paragraph Score Each paragraph p is then scored by combining its relevance to the headline and its contribution to covering high-quality, previously unselected clusters U:

$$\text{Score}^+(p) = \underbrace{\sum_{c \in \text{clusters}(p) \cap U} \text{ClusterScore}(c)}_{\text{Diversity Term}} + \lambda \cdot \underbrace{\text{Sim}(h, p)}_{\text{Relevance Term}}$$

The diversity term encourages the selection of paragraphs that introduce novel content by covering high-quality clusters that have not yet been selected, while the relevance term ensures alignment with the user query. The parameter λ controls the trade-off between diversity and relevance.

This enhanced version, denoted as **GreedyPlus**, leverages soft weighting to prioritize both informative and relevant content, resulting in greater alignment with the user's query while maintaining viewpoint diversity.

Efficiency Since Stage II operates on a small set of top-ranked candidate paragraphs (e.g., 100) from Stage I for fine-grained re-ranking, the additional cost of diversity-aware re-ranking remains relatively low. Our pipeline runs efficiently, with an average runtime of approximately 1.2 seconds per event (see Appendix B for details).

5 Data Construction

To evaluate diverse news retrieval, we construct two benchmark datasets: **LocalNews**, a new corpus of paragraph-annotated local news events, and **DS-Global**, an adaptation of the DiverseSumm dataset (Huang et al., 2024a) for retrieval. These datasets cover both local and global contexts and enable robust, fine-grained evaluation of diversity-aware retrieval systems.

5.1 Local News Benchmark (LocalNews)

LocalNews is built using Google News' "Full Coverage" feature to collect multi-source reports on

Properties	LocalNews	DSGlobal
# Events	103	147
# Paragraphs	5,296	7,532
Avg. Sentence Count	7.1	7.5
Avg. Word Count	124.9	123

Table 1: Data Statistics for LocalNews and DSGlobal.

the same event. We apply event de-duplication and segment articles into paragraphs capped at 512 tokens, preserving semantic boundaries. The final set includes 5,296 paragraphs averaging 7.1 sentences and 124.9 words each.

Each paragraph was labeled for relevance against a one-sentence abstractive event summary generated by GPT-4o-mini, which served as the query for simulating user retrieval intent. Summaries were iteratively refined up to five times if their associated relevance labels had low positive rates. Paragraphs were annotated in three rounds with majority voting (98.7% agreement), and those from outside the event's full coverage were auto-labeled as irrelevant.

5.2 Global News Benchmark (DSGlobal)

To assess generalizability beyond local news, we adapt DiverseSumm (Huang et al., 2024a), a dataset originally designed for news summarization, into a paragraph-level retrieval benchmark. The resulting **DSGlobal** dataset covers 147 global news events with 7,532 segmented paragraphs. Each paragraph contains 7.5 sentences and 123 words on average. Key statistics comparing LocalNews and DSGlobal are summarized in Table 1.

See Appendix A for full details on source diversity, data processing, annotation procedures, corresponding prompt templates used and paragraph distribution visualizations.

6 Evaluation

6.1 Experimental Setup

Given a news headline as the query, the system retrieves a ranked list of the top k paragraphs from the entire news corpus. Paragraphs unrelated to the event, either from irrelevant reports or different events, are treated as negatives. Stage I retrieves the top 100 relevant paragraphs as candidates for re-ranking. In Stage II, while the cluster coverage threshold offers a flexible stopping condition during greedy selection, we follow standard retrieval evaluation protocols and report fixed top-k results

Model	Relevancy			Diversity			Relevancy			Diversity		
	P	R	F1	D	I	C	P	R	F1	D	I	C
	top 5						**top 10**					
BM25	95.9	16.9	28.8	20.8	47.3	35.8	91.4	31.3	46.7	26.5	37	51.3
DenseRetr	**97.5**	**17.2**	**29.3**	17	45.2	34.2	**95**	**32.9**	**48.9**	23	36.5	52.9
MMR	93	16.6	28.1	29.8	54.3	41.5	92.8	32.4	48.1	29	38.8	56
DkMIPS	86.2	15	25.6	**30.9**	51.5	33.1	85	29.3	43.6	**32.5**	39.8	50.6
NEWSCOPE (GreedySCS)	95.3	16.8	28.6	24.2	64	52	93.7	32.6	48.3	26.3	44.7	62.2
NEWSCOPE (GreedyPlus)	92.2	16.1	27.5	29.8	**73.6**	**62.3**	90.3	31.7	46.9	30.2	**54.2**	**74.6**
	top 20						**top 50**					
BM25	77	49.6	60.3	37.7	28.5	69.7	46.3	66.5	54.6	**55.1**	17.5	83.2
DenseRetr	**85.7**	**56.1**	**67.8**	31.4	30.4	73.6	53.8	77	63.4	45.7	28	90.8
MMR	81.4	52.7	63.9	37.1	32.2	73.2	49.8	71.4	58.7	51.3	30.3	88.7
DkMIPS	76.6	50.4	60.8	**38.7**	33.4	69.1	51.8	75.3	61.4	49.8	29	88.5
NEWSCOPE (GreedySCS)	83.4	54.3	65.8	34.9	37.8	78.5	52.2	75	61.6	48.7	**38**	92.4
NEWSCOPE (GreedyPlus)	84.1	55.8	67.1	34.8	**38.8**	**84.9**	54.1	78	63.9	47.6	32.3	**92.7**

Table 2: Performance on **LocalNews** benchmark across relevance metrics (P, R, F1) and diversity metrics (D, I, C) at different retrieval depths of top 5, 10, 20, and 50 results.

with $k \in \{5, 10, 20, 50\}$. Paragraphs are ranked based on their selection order. The threshold remains tunable for customized coverage needs. λ is set to be 0.5 for all reported results.

6.2 Evaluation Metrics

We evaluate systems on both standard relevance metrics and novel metrics designed to capture fine-grained semantic diversity.

Precision (P), Recall (R), and F1-score (F1) Standard metrics for assessing the correctness and completeness of retrieved paragraphs.

Average Pairwise Distance (D) To assess internal diversity, we compute the average pairwise cosine distance among retrieved paragraph embeddings:

$$D = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} \left(1 - \cos(\mathbf{p}_i, \mathbf{p}_j)\right)$$

Higher values indicate greater semantic variation.

Positive Cluster Coverage (C) We define diversity in terms of sentence-level semantics. Each sentence is assigned to a semantic cluster (Section 4.2.1), and each cluster represents a unique perspective in news reporting. C measures the proportion of these clusters covered by the top-k retrieved paragraphs:

$$C = \frac{\text{\# Covered Clusters}}{\text{\# Total Clusters in Relevant Paragraphs}}$$

This metric captures diversity through fine-grained semantic recall and avoids rewarding redundancy.

Information Density Ratio (I) To evaluate informativeness, we compute the average number of unique clusters represented in the retrieved paragraphs:

$$I = \frac{\text{\# Covered Clusters}}{\text{\# Total Sentences in Retrieved Paragraphs}}$$

This penalizes repetition and favors concise, information-rich paragraphs.

Together, C, D, and I provide a multidimensional assessment of semantic coverage, viewpoint diversity, and content efficiency.

6.3 Baselines

We compare our framework against strong retrieval methods with different strategies for balancing relevance and diversity.

BM25 A sparse retrieval model based on term frequency and inverse document frequency (Robertson et al., 1995), highly effective for relevance matching but lacking diversity modeling.

DenseRetr (Dense Retrieval) A multilingual dense retriever `bilingual-embedding-large` (Thakur et al., 2020), fine-tuned on STS and NLI tasks. It is used in Stage I of our framework, strong on relevance but diversity-agnostic.

MMR (Maximal Marginal Relevance) A classical method that promotes diversity by selecting documents that maximize relevance to the query while minimizing redundancy with already selected items (Carbonell and Goldstein, 1998).

DkMIPS (Diversity-aware k-Maximum Inner Product Search) A method that jointly optimizes

Model	Relevancy			Diversity			Relevancy			Diversity		
	P	**R**	**F1**	**D**	**I**	**C**	**P**	**R**	**F1**	**D**	**I**	**C**
	top 5						**top 10**					
BM25	92.9	10.6	19	25.3	45.8	26	91.1	20.8	33.9	29.7	36.5	41.6
DenseRetr	**95.6**	**10.8**	**19.4**	20.7	45.3	25.1	**94.6**	**21.4**	**34.9**	24.6	36.2	40.9
MMR	90.9	10.3	18.4	**35.2**	51.9	30	90.9	20.6	33.6	32.9	40.8	46
D*k*MIPS	92.9	10.5	18.9	28.7	50.2	25.5	92.2	20.9	34	30.2	38.6	39.6
NEWSCOPE (GreedySCS)	92.7	10.5	18.9	29.4	66.1	43.5	92.2	20.9	34.1	29.7	46.8	53.8
NEWSCOPE (GreedyPlus)	91	10.3	18.5	34.6	**74.7**	**53.5**	87.6	19.8	32.2	**37.9**	**61.9**	**75**
	top 20						**top 50**					
BM25	84.9	38.4	52.9	37.4	28.6	61.5	57.3	61.6	59.4	55.3	18.7	81.9
DenseRetr	**93**	**42.1**	**57.9**	31.3	28.8	65.2	**72.8**	**77.6**	**75.2**	45.7	22.8	91.3
MMR	87.2	39.3	54.2	35.9	30.8	64.8	61.5	65.7	63.5	51.8	25.5	86.6
D*k*MIPS	90.1	40.6	56	34.9	29.9	62.2	70.4	75.5	72.9	48.2	23.8	89.8
NEWSCOPE (GreedySCS)	88.6	39.9	55	34.5	34	68.4	65.3	69.7	67.4	**49**	**31.1**	93.3
NEWSCOPE (GreedyPlus)	81.5	36.6	50.5	**40.3**	45.8	**88.5**	67.3	71.8	69.5	48.5	30.3	**94.9**

Table 3: Performance on **DSGlobal** benchmark across relevance metrics (P, R, F1) and diversity metrics (D, I, C) at different retrieval depths of top 5, 10, 20, and 50 results.

Rank	DenseRetr	NEWSCOPE
1	Altman expresses excitement and concern.	Altman: benefits outweigh risks.
2	Altman: benefits outweigh risks.	Russia's AI interest; hallucinations.
3	Altman expresses excitement and concern.	ChatGPT adoption and benchmarks.
4	Altman: benefits outweigh risks.	Altman vs. Virginia Governor on AI in education.
5	Altman expresses excitement and concern.	Criticism: ChatGPT unreliable and uncreative.

Table 4: Top-5 retrieved paragraphs (summarized) under dense retrieval and proposed diverse retrieval.

(a) LocalNews

(b) DSGlobal

Figure 3: Visualization of model performance across different retrieval depth levels.

for relevance and diversity by maximizing similarity to the query and minimizing pairwise similarity among retrieved items (Huang et al., 2024b).

All methods use the same input query (event headline) and top-k output size. The formal scoring functions for baselines BM25, MMR, and D*k*MIPS are provided in Appendix C.

7 Results & Analysis

7.1 Main Results

We report results on the LocalNews and DSGlobal benchmarks in Table 2 and Table 3, evaluated across four retrieval depths (top 5, 10, 20, and 50) using standard relevance metrics, Precision (P), Recall (R), and F1, and diversity-oriented metrics, Average Pairwise Distance (D), Information Density Ratio (I), and Positive Cluster Coverage (C).

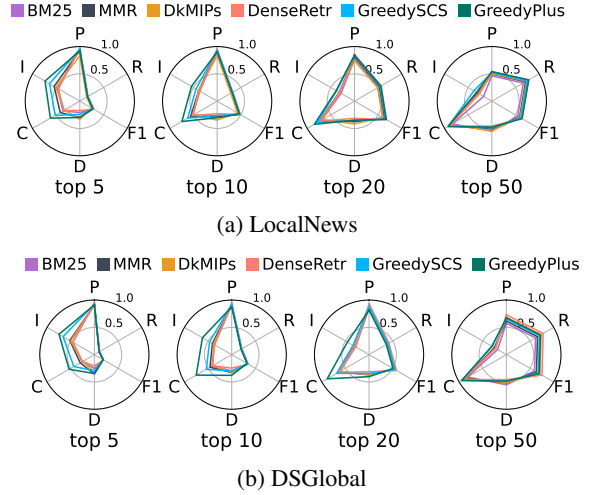**Relevance-focused baselines.** Among retrieval methods, BM25 and DenseRetr achieve strong relevance scores. DenseRetr yields the highest precision, recall, and F1, demonstrating effective semantic matching. However, it consistently underperforms on diversity metrics (D, I, and C), reflecting its tendency to retrieve redundant content with limited perspective variation.

Diversity-aware baselines. MMR and D*k*MIPS improve diversity through novelty-aware selection. MMR yields modest gains in diversity, especially at shallow depths. D*k*MIPS achieves stronger improvements in diversity, but sacrifices relevance performance, leading to lower overall F1.

Effectiveness of NEWSCOPE. As shown in Figure 3, both variants of **NEWSCOPE** substan-

7

Model	LocalNews															
	Top 5				Top 10				Top 20				Top 50			
	F1	D	I	C	F1	D	I	C	F1	D	I	C	F1	D	I	C
GreedyPlus	27.5	29.8	62.3	73.6	46.9	30.2	74.6	54.2	67.1	34.8	84.9	38.8	63.9	47.6	92.7	32.3
w/o diversity term	**29.3**	17.7	35.8	45.2	**48.8**	24.1	52.9	36.1	**68.3**	33.3	75.1	30.1	**64.4**	47.2	90.9	26.7
w/o relevance term	25.2	**34.4**	**64.6**	**82.3**	32.0	**42.9**	**80.0**	**76.0**	30.6	**51.0**	**87.0**	**72.3**	57.4	**49.2**	**93.1**	**42.7**
DSGlobal																
GreedyPlus	18.5	34.6	74.7	53.5	32.2	37.9	61.9	75.0	50.5	40.3	45.8	88.5	69.5	48.5	30.3	**94.9**
w/o diversity term	**19.4**	22.2	46.6	25.4	**34.6**	26.9	37.1	40.8	**56.3**	33.5	29.8	64.9	**74.5**	46.7	22.8	91.4
w/o relevance term	18.5	**35.2**	**76.8**	**53.6**	30.3	**41.5**	**66.6**	**76.0**	35.5	**51.3**	**58.4**	**89.3**	59.1	**52.3**	**36.0**	94.8

Table 5: Ablation study of **NEWSCOPE (GreedyPlus)** on **LocalNews** and **DSGlobal.**

tially outperform baselines on diversity metrics. GreedySCS achieves a strong balance, enhancing Positive Cluster Coverage (C) with minimal precision loss. GreedyPlus further boosts diversity, attaining the highest scores on Information Density Ratio (I) and C across nearly all settings. We observe greater gains at deeper retrieval depths (top 20 and 50). Performance trends are consistent across LocalNews and DSGlobal, confirming the robustness of our framework.

Depth-wise trade-offs. Figure 3 illustrates how diversity improves with retrieval depth. While most methods perform similarly at top 5 and 10, diversity metrics (D, I, C) increase significantly for GreedySCS and GreedyPlus at top 20 and 50, without compromising relevance. This highlights their effectiveness in mitigating redundancy and uncovering broader semantic coverage.

Overall, the results demonstrate that **NEWSCOPE** achieves the best balance of relevance and diversity, enabling more comprehensive event understanding.

7.2 Qualitative Analysis

We compare relevance-based retrieval (DenseRetr) with our diversity-aware re-ranking (GreedyPlus) to illustrate improvements in factual coverage and reduction in redundancy.

As shown in Table 4, Given the query "*OpenAI CEO Sam Altman expresses concerns about AI's risks and highlights the need for careful regulation amidst competition from global players*", DenseRetr returns near-duplicate content focused on Altman's interview, while GreedyPlus retrieves a wider range of perspectives, including global competition, societal risks, policy debates, and criticisms of generative AI. This illustrates how sentence-level modeling and cluster-based re-ranking enable broader yet relevant coverage. Full examples appear in Appendix D.

7.3 Ablation Study

We conduct an ablation study to assess the contributions of the relevance and diversity components in GreedyPlus. Table 5 reports summarized results on both datasets; full tables are provided in Appendix E.1.

Removing the diversity term reduces the method to a relevance-based reranking. While slightly improving F1, it significantly reduces all diversity metrics, leading to redundant results. In contrast, removing the relevance term maximizes diversity but harms F1, selecting loosely relevant or off-topic content. GreedyPlus consistently achieves the best balance across retrieval depths, demonstrating the importance of jointly modeling relevance and diversity. Qualitative examples for the ablation study are provided in Appendix E.2.

8 Conclusion

In this work, we propose **NEWSCOPE**, a two-stage framework for diverse news retrieval that enhances event coverage by modeling semantic variation with fine-grained granularity. By integrating sentence-level clustering with interpretable greedy re-ranking, our method surfaces complementary perspectives without sacrificing relevance. We propose three novel diversity metrics, construct two benchmarks spanning local and global contexts, and show consistent gains over strong baselines. By revealing underrepresented viewpoints, this work highlights the importance of diversity-aware retrieval in mitigating redundancy and promoting unbiased access to information. Future directions include domain generalization and integration with fact verification for more trustworthy information retrieval.

8

Limitations

While our approach improves diversity in news retrieval, several limitations remain. First, although we focus on the news domain where divergent perspectives naturally occur and diversity is especially critical, our framework is domain-agnostic and can be applied to other areas. Future work may investigate its adaptability to domains such as finance, science, or law, where diversity takes different forms and factual consistency is often prioritized over viewpoint variation.

Second, our selection mechanism optimizes for semantic variation at the sentence level but does not explicitly control for factual consistency across retrieved paragraphs. As a result, the retrieved set may contain conflicting or misleading narratives. Future extensions could incorporate source credibility or factual verification to better balance diversity, accuracy, and contextual coherence in the retrieval output.

References

Shunit Agmon, Amir Gilad, Brit Youngmann, Shahar Zoarets, and Benny Kimelfeld. 2024. Finding convincing views to endorse a claim. *CoRR*, abs/2408.14974.

Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336.

Laming Chen, Guoxin Zhang, and Eric Zhou. 2018. Fast greedy MAP inference for determinantal point process to improve recommendation diversity. In *NeurIPS*, pages 5627–5638.

Sihao Chen, Daniel Khashabi, Chris Callison-Burch, and Dan Roth. 2019. Perspectroscope: A window to the world of diverse perspectives. In *ACL (3)*, pages 129–134. Association for Computational Linguistics.

Ting-Chih Chen, Chia-Wei Tang, and Chris Thomas. 2024. Metasumperceiver: Multimodal multi-document evidence summarization for fact-checking. In *ACL*, pages 8742–8757.

Yuwei Chuai, Anastasia Sergeeva, Gabriele Lenzini, and Nicolas Pröllochs. 2024. Community fact-checks trigger moral outrage in replies to misleading posts on social media. *CoRR*, abs/2409.08829.

Andrew Estornell, Sanmay Das, and Yevgeniy Vorobeychik. 2020. Deception through half-truths. In *AAAI*, pages 10110–10117.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *ACL*, pages 1074–1084.

Mohsen Fayyaz, Ali Modarressi, Hinrich Schuetze, and Nanyun Peng. 2025. Collapse of dense retrievers: Short, early, and literal biases outranking factual evidence. *arXiv preprint arXiv:2503.05037*.

Yi R. Fung, Kung-Hsiang Huang, Preslav Nakov, and Heng Ji. 2022. The battlefront of combating misinformation and coping with media bias. In *KDD*, pages 4790–4791.

Xuming Hu, Zhijiang Guo, Guanyu Wu, Lijie Wen, and Philip S. Yu. 2023. Give me more details: Improving fact-checking with latent retrieval. *CoRR*, abs/2305.16128.

Kung-Hsiang Huang, Philippe Laban, Alexander R. Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024a. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *NAACL-HLT*, pages 570–593.

Qiang Huang, Yanhao Wang, Yiqun Sun, and Anthony Kum Hoe Tung. 2024b. Diversity-aware k-maximum inner product search revisited. *CoRR*, abs/2402.13858.

Israa Jaradat, Haiqi Zhang, and Chengkai Li. 2024. On detecting cherry-picking in news coverage using large language models. *CoRR*, abs/2401.05650.

Zhengbao Jiang, Ji-Rong Wen, Zhicheng Dou, Wayne Xin Zhao, Jian-Yun Nie, and Ming Yue. 2017. Learning to diversify search results via subtopic attention. In *SIGIR*, pages 545–554. ACM.

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.

Brendan King and Jeffrey Flanigan. 2023. Diverse retrieval-augmented in-context learning for dialogue state tracking. In *ACL*, pages 5570–5585.

Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2019. Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal*, 22:188–227.

9

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2014–2037.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *NIST Special Publication*, (500225):109–123.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Syst. Appl.*, 237(Part C):121641.

Daniel Russo, Stefano Menini, Jacopo Staiano, and Marco Guerini. 2024. Face the facts! evaluating rag-based fact-checking pipelines in realistic settings. *CoRR*, abs/2412.15189.

Sandeep Singamsetty, Nishtha Madaan, Sameep Mehta, Varad Bhatnagar, and Pushpak Bhattacharyya. 2023. "beware of deception": Detecting half-truth and debunking it through controlled claim editing. *CoRR*, abs/2308.07973.

Ronit Singhal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using rag and few-shot in-context learning with llms. *CoRR*, abs/2408.12060.

Timo Spinde, Christin Jeggle, Magdalena Haupt, Wolfgang Gaissmaier, and Helge Giese. 2022. How do we raise media bias awareness effectively? effects of visualizations to communicate bias. *Plos one*, 17(4):e0266204.

Aniruddh Sriram, Fangyuan Xu, Eunsol Choi, and Greg Durrett. 2024. Contrastive learning to improve retrieval for real-world fact checking. *CoRR*, abs/2410.04657.

Yiqun Sun, Qiang Huang, Yanhao Wang, and Anthony K. H. Tung. 2024. Diversinews: Enriching news consumption with relevant yet diverse news articles retrieval. *Proc. VLDB Endow.*, 17(12):4277–4280.

Megha Sundriyal, Harshit Choudhary, Tanmoy Chakraborty, and Md. Shad Akhtar. 2024. Crowd intelligence for early misinformation prediction on social media. *CoRR*, abs/2408.04463.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2020. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv e-prints*, pages arXiv–2010.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: a heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*, pages 809–819.

Zhengjia Wang, Danding Wang, Qiang Sheng, Juan Cao, Silong Su, Yifan Sun, Beizhe Hu, and Siyuan Ma. 2023. Understanding news creation intents: Frame, dataset, and method. *CoRR*, abs/2312.16490.

Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2016. Modeling document novelty with neural tensor network for search result diversification. In *SIGIR*, pages 395–404. ACM.

Jun Yu, Sunil Mohan, Duangmanee Putthividhya, and Weng-Keen Wong. 2014. Latent dirichlet allocation based diversified retrieval for e-commerce search. In *WSDM*, pages 463–472. ACM.

Xiaoying Zhang, Hongning Wang, and Hang Li. 2023. Disentangled representation for diversified recommendations. In *WSDM*, pages 490–498. ACM.

Yusen Zhang, Nan Zhang, Yixin Liu, Alexander R. Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen R. McKeown, and Rui Zhang. 2024. Fair abstractive summarization of diverse perspectives. In *NAACL-HLT*, pages 3404–3426.

10

A Data Construction Details: Source Diversity and Annotation Process

A.1 Data Sources

LocalNews is curated to support sentence-level diversity modeling in event-centric retrieval. It focuses on localized events and is constructed using Google News' "Full Coverage" feature, which aggregates articles from various sources reporting on the same event. This method helps mitigate source bias by incorporating a broad spectrum of viewpoints.

We analyzed the collected sources and found a mix of:

- **Mainstream outlets** (e.g., major national newspapers and broadcasters),

- **Independent media platforms**,

- **Domain-specific publishers** (e.g., business- or health-focused),

- **Regional and international media**, including nearby countries and global organizations.

To ensure consistency and avoid personalization bias, we accessed "Full Coverage" pages from varying IP addresses (via VPN), devices, login states (logged-in vs. incognito), and locations. The resulting content remained stable across conditions.

A.2 Data Processing

We applied the following preprocessing steps to construct the benchmark:

- **Event De-duplication**: Articles reporting on the same incident were grouped and deduplicated to ensure a clean set of distinct events.

- **Paragraph Segmentation**: Articles were segmented into coherent paragraphs, capped at 512 tokens, which aligns with the input limits of modern encoding models. Paragraph boundaries are preserved wherever possible; overly long paragraphs are truncated at sentence boundaries. The final dataset contains 5,296 paragraphs, averaging 7.1 sentences and 124.9 words each.

A.3 Annotation Procedure

We adopt an event-centric labeling strategy:

- **Query Generation**: One-sentence abstractive summaries were generated using GPT-4o-mini to to represent each event. These summaries served as queries for retrieval, simulating user searches for event-related information. To ensure high-quality queries, we iteratively refined summaries where the positive rate of relevance labels did not meet a predefined threshold with up to five rounds.

- **Relevance Labeling**: Each paragraph was annotated for relevance to the event summary. We used majority voting across three rounds of annotation, achieving an inter-annotation agreement rate of 98.7%. Paragraphs from articles outside the event's full coverage were auto-labeled as irrelevant.

A.4 Prompt Templates

We employ LLM prompting to assist in dataset labelling for (1) event headline generation and (2) paragraph relevance annotation. Below are the two prompt templates:

Prompt: Generate Summarized Headline

Read the following news articles discussing the same event. Produce a very brief headline that summarizes the general aspects of the event **in one sentence**. Include names of key entities (e.g., people, countries, companies, or products) when relevant to help refer to the event. Only output the headline, without any explanation.
==========
Article 0: {content of Article 0}
Article 1: {content of Article 1}
Article 2: {content of Article 2}
... (repeat until max_token is reached)
==========
Reminder: Your answer must be a one-sentence headline focusing on the main event and important entities. Do not include additional explanation.

A.5 Paragraph Distribution

Figure 4 compares the distribution of positive and negative paragraphs per event, while Figure 5 shows the variation in the number of positive paragraphs across events in both datasets.
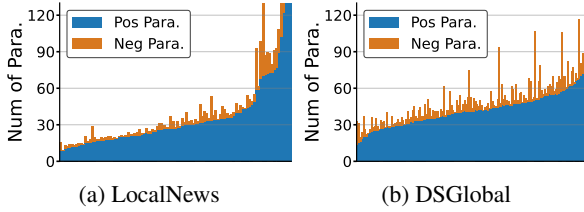

(a) LocalNews   (b) DSGlobal

Figure 4: Positive vs. negative paragraph counts per event.
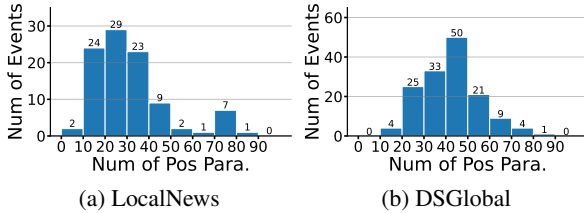

(a) LocalNews   (b) DSGlobal

Figure 5: Event distribution by number of positive paragraphs.

B Efficiency Analysis

While our framework introduces a second-stage re-ranking module, it is designed with practical efficiency in mind. In Stage I, we retrieve the top 100 paragraphs using a dense retriever, which acts as a lightweight filter. This greatly reduces the candidate pool for Stage II, where sentence-level clustering and re-ranking are applied.

Retrieving 50 final items from this pool already covers over 80% of relevant clusters on average (as shown in main results), confirming that 100 candidates provide sufficient scope for effective and diverse re-ranking.

We benchmarked the average runtime across 100 events to assess the end-to-end efficiency of the **NEWSCOPE** pipeline:

Step	Time (s)
Dense retrieval	0.025
Sentence encoding & clustering	0.84
Re-ranking (Greedy Selection)	0.51
Total	**1.175**

Table 6: Average runtime per event across 100 examples.

This breakdown demonstrates that our full pipeline operates within 1.2 seconds per event, confirming its scalability for real-world diverse news retrieval applications.

C Baseline Scoring Functions

BM25 BM25 (Robertson et al., 1995) ranks documents based on term frequency and inverse document frequency, balancing relevance with length normalization. The scoring function is:

$$\text{BM25}(Q, D) = \sum_{t \in Q} \log \frac{N - n_t + 0.5}{n_t + 0.5}$$
$$\cdot \frac{(k_1 + 1) f_{t,D}}{k_1 (1 - b + b \cdot \frac{|D|}{\text{avgdl}}) + f_{t,D}}$$

MMR (Maximal Marginal Relevance) MMR (Carbonell and Goldstein, 1998) balances relevance and diversity by selecting the document D_i that maximizes:

$$\text{MMR}(D_i, Q) = \arg \max_{D_i \in \mathcal{D} \setminus S} \Big(\lambda \cdot \text{Sim}(D_i, Q)$$
$$- (1 - \lambda) \cdot \max_{D_j \in S} \text{Sim}(D_i, D_j) \Big)$$

where $\lambda \in [0, 1]$ controls the trade-off.

DkMIPS DkMIPS (Huang et al., 2024b) formulates diversity-aware retrieval as a penalized inner product optimization:

$$f(S) := \lambda \sum_{\mathbf{p} \in S} \langle \mathbf{p}, \mathbf{q} \rangle - \frac{2\mu(1 - \lambda)}{k(k - 1)} \sum_{\mathbf{p}, \mathbf{p}' \in S} \langle \mathbf{p}, \mathbf{p}' \rangle$$

This encourages both high relevance to the query $\mathbf{q}$ and low redundancy among retrieved paragraphs $\mathbf{p} \in S$.

D   Qualitative Analysis Example: Full Text

The following table contains the retrieval results of news headline: "*OpenAI CEO Sam Altman expresses concerns about AI's risks while highlighting ChatGPT's transformative potential and the need for careful regulation amidst competition from global players like China and Russia.*"

These results illustrate how our method surfaces complementary angles, including technological, political, educational, and critical, enhancing users' understanding of complex events beyond repetitive core quotes.

Table 7: Comparison of Top-5 Results from Relevance Retrieval and Diverse Retrieval

Relevance Retrieval: DenseRetr	Diverse Retrieval: NEWSCOPE (GreedyPlus)
1 Sam Altman, co-founder and chief executive officer of OpenAI Inc., speaks during TechCrunch Disrupt 2019 in San Francisco, California, on Thursday, Oct. 3, 2019.OpenAI CEO Sam Altman said in a recent interview with ABC News that he's a "little bit scared" of artificial intelligence technology and how it could affect the workforce, elections and the spread of disinformation.OpenAI developed the ChatGPT bot, which creates human-like answers to questions and ignited a new AI craze."I think people really have fun with [ChatGPT]," Altman said in the interview.But his excitement over the transformative potential of AI technology, which Altman said will eventually reflect "the collective power, and creativity, and will of humanity," was balanced by his concerns about "authoritarian regimes" developing competing AI technology."We do worry a lot about authoritarian governments developing this," Altman said. Overseas governments have already begun to bring competing AI technology to market.	Among the concerns of the destructive capabilities of this technology is the replacement of jobs. Altman says this will likely replace some jobs in the near future, and worries how quickly that could happen."I think over a couple of generations, humanity has proven that it can adapt wonderfully to major technological shifts," Altman said. "But if this happens in a single-digit number of years, some of these shifts ... That is the part I worry about the most."But he encourages people to look at ChatGPT as more of a tool, not as a replacement. He added that "human creativity is limitless, and we find new jobs. We find new things to do."ABC News OpenAI CEO Sam Altman speaks with ABC News, Mar. 15, 2023.The ways ChatGPT can be used as tools for humanity outweigh the risks, according to Altman."We can all have an incredible educator in our pocket that's customized for us, that helps us learn," Altman said. "We can have medical advice for everybody that is beyond what we can get today."
2 Among the concerns of the destructive capabilities of this technology is the replacement of jobs. Altman says this will likely replace some jobs in the near future, and worries how quickly that could happen."I think over a couple of generations, humanity has proven that it can adapt wonderfully to major technological shifts," Altman said. "But if this happens in a single-digit number of years, some of these shifts ... That is the part I worry about the most."But he encourages people to look at ChatGPT as more of a tool, not as a replacement. He added that "human creativity is limitless, and we find new jobs. We find new things to do."ABC News OpenAI CEO Sam Altman speaks with ABC News, Mar. 15, 2023.The ways ChatGPT can be used as tools for humanity outweigh the risks, according to Altman."We can all have an incredible educator in our pocket that's customized for us, that helps us learn," Altman said. "We can have medical advice for everybody that is beyond what we can get today."	Countries such as Russia have shown an interest in pursuing AI. Russian President Vladimir Putin told students in 2017 that whoever led the AI race would most likely "rule the world." GPT-4 has safeguards to protect users from engaging in illicit conduct, such as seeking information about how to construct bombs.Altman also noted the software's inability to fact-check. "The thing that I try to caution people the most is what we call the 'hallucinations problem,'" he said. "The model will confidently state things as if they were facts that are entirely made up." OpenAI is attempting to counter this problem by having the bot use deductive reasoning rather than memorization, allowing it to process statements in real time.OpenAI launched the latest version of its software, GPT-4, on Tuesday. The bot has a faster response rate and can process image prompts.CLICK HERE TO READ MORE FROM THE WASHINGTON EXAMINER
3 ABC News(NEW YORK) – The CEO behind the company that created ChatGPT believes artificial intelligence technology will reshape society as we know it. He believes it comes with real dangers, but can also be "the greatest technology humanity has yet developed" to drastically improve our lives."We've got to be careful here," said Sam Altman, CEO of OpenAI. "I think people should be happy that we are a little bit scared of this."Altman sat down for an exclusive interview with ABC News' chief business, technology and economics correspondent Rebecca Jarvis to talk about the rollout of GPT-4 – the latest iteration of the AI language model.In his interview, Altman was emphatic that OpenAI needs both regulators and society to be as involved as possible with the rollout of ChatGPT – insisting that feedback will help deter the potential negative consequences the technology could have on humanity. He added that he is in "regular contact" with government officials.	ChatGPT is an AI language model, the GPT stands for Generative Pre-trained Transformer.Released only a few months ago, it is already considered the fastest-growing consumer application in history. The app hit 100 million monthly active users in just a few months. In comparison, TikTok took nine months to reach that many users and Instagram took nearly three years, according to a UBS study.Watch the exclusive interview with Sam Altman on "World News Tonight with David Muir" at 6:30 p.m. ET on ABC.Though "not perfect," per Altman, GPT-4 scored in the 90th percentile on the Uniform Bar Exam. It also scored a near-perfect score on the SAT Math test, and it can now proficiently write computer code in most programming languages.GPT-4 is just one step toward OpenAI's goal to eventually build Artificial General Intelligence, which is when AI crosses a powerful threshold which could be described as AI systems that are generally smarter than humans.

Relevance Retrieval: DenseRetr	Diverse Retrieval: NEWSCOPE (GreedyPlus)
4 Among the concerns of the destructive capabilities of this technology is the replacement of jobs. Altman says this will likely replace some jobs in the near future, and worries how quickly that could happen."I think over a couple of generations, humanity has proven that it can adapt wonderfully to major technological shifts," Altman said. "But if this happens in a single-digit number of years, some of these shifts ... That is the part I worry about the most."But he encourages people to look at ChatGPT as more of a tool, not as a replacement. He added that "human creativity is limitless, and we find new jobs. We find new things to do."OpenAI CEO Sam Altman speaks with ABC News, Mar. 15, 2023. ABC NewsThe ways ChatGPT can be used as tools for humanity outweigh the risks, according to Altman."We can all have an incredible educator in our pocket that's customized for us, that helps us learn," Altman said. "We can have medical advice for everybody that is beyond what we can get today."	He continued: "It is going to eliminate a lot of current jobs, that's true. We can make much better ones. The reason to develop AI at all, in terms of impact on our lives and improving our lives and upside, this will be the greatest technology humanity has yet developed."Altman also spoke on the impacts that AI-powered chatbots would have on education and whether it would "increase laziness among students.""Education is going to have to change," the OpenAI CEO said. "But it's happened many other times with technology. When we got the calculator, the way we taught math and what we tested students on totally changed."VIRGINIA GOV. YOUNGKIN SAYS MORE SCHOOLS SHOULD BAN CHATGPTVirginia Gov. Glenn Youngkin announced in March that more school districts should ban AI technologies like ChatGPT.Youngkin said that the goal of education was "to make sure that our kids can think and, therefore, if a machine is thinking for them, then we're not accomplishing our goal."
5 OpenAI CEO Sam Altman believes artificial intelligence has incredible upside for society, but he also worries about how bad actors will use the technology.In an ABC News interview this week, he warned "there will be other people who don't put some of the safety limits that we put on."OpenAI released its A.I. chatbot ChatGPT to the public in late November, and this week it unveiled a more capable successor called GPT-4.Other companies are racing to offer ChatGPT-like tools, giving OpenAI plenty of competition to worry about, despite the advantage of having Microsoft as a big investor."It's competitive out there," OpenAI cofounder and chief scientist Ilya Sutskever told The Verge in an interview published this week. "GPT-4 is not easy to develop...there are many many companies who want to do the same thing, so from a competitive side, you can see this as a maturation of the field."	Chat GPT is an attempt to make AI into a conversation between humans and the technology via the computer. People ask it questions and get text message-style answers. Regarding Chat GPT, Marks dismissed it as unreliable: "They don't tell the truth. In fact, Chat GPT, when you log on to it, says... 'Don't trust the facts that we're telling you.'" (The full chat GPT warning reads: "While we have safeguards in place, the system may occasionally generate incorrect or misleading information and produce offensive or biased content. It is not intended to give advice.")AI EXPERTS WEIGH DANGERS, BENEFITS OF CHATGPT ON HUMANS, JOBS AND INFORMATION: 'DYSTOPIAN WORLD'Marks said of Chat GPT-style software: "They aren't creative. They don't understand. They have a terrible sense of humor."

E Ablation Study

E.1 Full Results

We present the complete ablation results for **NEWSCOPE (GreedyPlus)** on **LocalNews** (Table 8) and **DSGlobal** (Table 9). **Removing the diversity term** improves F1 slightly but leads to sharp drops in D, I, and C, indicating increased redundancy and weaker perspective coverage. **Removing the relevance term** boosts diversity (with top scores in D, I, and C) but significantly lowers F1, suggesting inclusion of less relevant content. These findings reinforce that both relevance and diversity are essential.

Model	P	R	F1	D	I	C	P	R	F1	D	I	C
			top 5						top 10			
NEWSCOPE (GreedyPlus)	92.2	16.1	27.5	29.8	62.3	73.6	90.3	31.7	46.9	30.2	74.6	54.2
w/o diversity term	**97.3**	**17.2**	**29.3**	17.7	35.8	45.2	**94.6**	**32.9**	**48.8**	24.1	52.9	36.1
w/o relevance term	87.6	14.7	25.2	**34.4**	**64.6**	**82.3**	68.2	20.9	32.0	**42.9**	**80.0**	**76.0**
			top 20						top 50			
NEWSCOPE (GreedyPlus)	84.1	55.8	67.1	34.8	84.9	38.8	54.1	78.0	63.9	47.6	92.7	32.3
w/o diversity term	**86.1**	**56.6**	**68.3**	33.3	75.1	30.1	**54.5**	**78.6**	**64.4**	47.2	90.9	26.7
w/o relevance term	43.1	23.7	30.6	**51.0**	**87.0**	**72.3**	48.5	70.3	57.4	**49.2**	**93.1**	**42.7**

Table 8: Ablation study of NEWSCOPE (GreedyPlus) on the **LocalNews** benchmark.

Model	P	R	F1	D	I	C	P	R	F1	D	I	C
			top 5						top 10			
NEWSCOPE (GreedyPlus)	91.0	10.3	18.5	34.6	74.7	53.5	87.6	19.8	32.2	37.9	61.9	75.0
w/o diversity term	**95.8**	**10.8**	**19.4**	22.2	46.6	25.4	**93.6**	**21.2**	**34.6**	26.9	37.1	40.8
w/o relevance term	90.9	10.3	18.5	**35.2**	**76.8**	**53.6**	83.7	18.5	30.3	**41.5**	**66.6**	**76.0**
			top 20						top 50			
NEWSCOPE (GreedyPlus)	81.5	36.6	50.5	40.3	45.8	88.5	67.3	71.8	69.5	48.5	30.3	**94.9**
w/o diversity term	**90.5**	**40.9**	**56.3**	33.5	29.8	64.9	**72.1**	**77.0**	**74.5**	46.7	22.8	91.4
w/o relevance term	60.2	25.2	35.5	**51.3**	**58.4**	**89.3**	56.9	61.4	59.1	**52.3**	**36.0**	94.8

Table 9: Ablation study of NEWSCOPE (GreedyPlus) on the **DSGlobal** benchmark.

E.2 Qualitative Analysis for Ablation Study

To further support the findings in Section 6.3, we provide a qualitative comparison of top-20 retrieved results for each ablation variant under the DSGlobal benchmark.

NEWSCOPE(GreedyPlus). Retrieves content spanning multiple dimensions, including Altman's interviews, global reactions, adoption benchmarks, and criticism, offering comprehensive coverage.

w/o Diversity Term. This reduces to a relevance-only model. The retrieved paragraphs are repetitive, with most centered around similar quotes from Altman. This improves precision but fails to uncover distinct viewpoints.

w/o Relevance Term. This variant retrieves many semantically distinct paragraphs, including out-of-context or weakly related content (e.g., speculative commentary, minor international opinions). While diverse, it lacks cohesion and topical alignment.

These patterns are reflected quantitatively in the main paper and qualitatively here, reinforcing that both scoring components are necessary to balance informativeness and perspective diversity.

Table 10: Example: Top-20 Paragraphs Retrieved by Three Strategies

Rank	w/o diversity term	NEWSCOPE (GreedyPlus)	w/o relevance term
1	Archaeologists discovered the oldest pearling town in the Persian Gulf on Siniyah Island in Umm al-Quwain, dating back to the late 6th century.	**Ranked 1 by sim_score.** Archaeologists discovered the oldest pearling town in the Persian Gulf on Siniyah Island in Umm al-Quwain, dating back to the late 6th century.	**Ranked 1 by sim_score.** Archaeologists discovered the oldest pearling town in the Persian Gulf on Siniyah Island in Umm al-Quwain, dating back to the late 6th century.
2	Timothy Power claimed that this is the oldest example of that kind of very specifically Khaleeji pearling town.	**Ranked 21 by sim_score.** Houses in the pearling town are densely packed. Umm al-Quwain plans to build a visitor's center at the site.	**Ranked 21 by sim_score.** Houses in the pearling town are densely packed. Umm al-Quwain plans to build a visitor's center at the site.
3	same as above	**Ranked 14 by sim_score.** The pearling town on Siniyah Island, near an ancient Christian monastery, spans 12 hectares and shows evidence of year-round habitation and social stratification.	**Ranked 14 by sim_score.** The pearling town on Siniyah Island, near an ancient Christian monastery, spans 12 hectares and shows evidence of year-round habitation and social stratification.
4	same as above	**Ranked 11 by sim_score.** Archaeologists discovered a 1,300-year-old pearling town on Siniya Island in Umm Al Quwain, marking the oldest such site in the Arabian Gulf with year-round habitation.	**Ranked 11 by sim_score.** Archaeologists discovered a 1,300-year-old pearling town on Siniya Island in Umm Al Quwain, marking the oldest such site in the Arabian Gulf with year-round habitation.
5	same as above	**Ranked 2 by sim_score.** Timothy Power claimed that this is the oldest example of that kind of very specifically Khaleeji pearling town.	**Ranked 2 by sim_score.** Timothy Power claimed that this is the oldest example of that kind of very specifically Khaleeji pearling town.
6	same as above	**Ranked 12 by sim_score.** The oldest known Khaleeji pearling town, found on Siniyah Island in Umm al-Quwain, predates Dubai and is near an ancient Christian monastery.	**Ranked 12 by sim_score.** The oldest known Khaleeji pearling town, found on Siniyah Island in Umm al-Quwain, predates Dubai and is near an ancient Christian monastery.
7	same as above	**Ranked 31 by sim_score.** The collapse of pearling after WWI mirrors the UAE's challenge of transitioning from fossil fuels to a carbon-neutral future, as highlighted by discoveries of oyster shell remains on Siniya Island.	**Ranked 40 by sim_score.** Multiple organizations excavated the ancient pearling town, with plans for a visitor's center amid Umm al-Quwain's efforts to balance development, including a $675 million project, with lessons from its historical sites.
8	same as above	**Ranked 28 by sim_score.** The pearling town on Siniya Island, which protects the Khor Al Beida marshlands in Umm Al Quwain, features year-round habitation and artifacts from the pearling industry, alongside a 1,400-year-old Christian monastery.	**Ranked 28 by sim_score.** The pearling town on Siniya Island, which protects the Khor Al Beida marshlands in Umm Al Quwain, features year-round habitation and artifacts from the pearling industry, alongside a 1,400-year-old Christian monastery.
9	Archaeologists discovered the oldest pearling town in the Persian Gulf on Siniyah Island in Umm al-Quwain, UAE, dating back to the late 6th century.	**Ranked 30 by sim_score.** The pearling town, spanning 12 hectares south of an ancient monastery, features diverse homes indicating social stratification and year-round habitation, with artifacts like pearls and diving weights found inside.	**Ranked 10 by sim_score.** The discovery of a 12-hectare pearling town on Al Siniya Island, dating to the pre-Islamic era, highlights Umm Al Quwain's thriving pearl trade and historical significance.
10	The discovery of a 12-hectare pearling town on Al Siniya Island, dating to the pre-Islamic era, highlights Umm Al Quwain's thriving pearl trade and historical significance.	**Ranked 40 by sim_score.** Multiple organizations excavated the ancient pearling town, with plans for a visitor's center amid Umm al-Quwain's efforts to balance development, including a $675 million project, with lessons from its historical sites.	**Ranked 31 by sim_score.** The collapse of pearling after WWI mirrors the UAE's challenge of transitioning from fossil fuels to a carbon-neutral future, as highlighted by discoveries of oyster shell remains on Siniya Island.

Table continued in next page.

Rank	w/o diversity term	NEWSCOPE (GreedyPlus)	w/o relevance term
11	Archaeologists discovered a 1,300-year-old pearling town on Siniya Island in Umm Al Quwain, marking the oldest such site in the Arabian Gulf with year-round habitation.	**Ranked 10 by sim_score.** The discovery of a 12-hectare pearling town on Al Siniya Island, dating to the pre-Islamic era, highlights Umm Al Quwain's thriving pearl trade and historical significance.	**Ranked 44 by sim_score.** Umm al-Quwain plans a $675 million real estate development including a bridge to Siniyah Island, hoping to boost the economy while learning from ancient pearling sites.
12	The oldest known Khaleeji pearling town, found on Siniyah Island in Umm al-Quwain, pre-dates Dubai and is near an ancient Christian monastery.	**Ranked 2 by sim_score.**	**irrelevant**
13	**same as above**	**Ranked 2 by sim_score.**	**irrelevant**
14	The pearling town on Siniyah Island, near an ancient Christian monastery, spans 12 hectares and shows evidence of year-round habitation and social stratification.	**Ranked 2 by sim_score.**	**irrelevant**
15	**same as above**	**Ranked 2 by sim_score.**	**irrelevant**
16	**same as above**	**Ranked 2 by sim_score.**	**irrelevant**
17	**same as above**	**Ranked 12 by sim_score.**	**irrelevant**
18	**same as above**	**Ranked 14 by sim_score.**	**irrelevant**
19	**same as above**	**Ranked 14 by sim_score.**	**irrelevant**
20	**same as above**	**Ranked 14 by sim_score.**	**irrelevant**