Open-World Evaluation for Retrieving Diverse Perspectives

Anonymous ACL submission

Abstract

We study retrieving a set of documents that covers various perspectives on a complex and contentious question (e.g., will ChatGPT do more harm than good?). We curate a **Be**nchmark for **R**etrieval **D**iversity for **S**ubjective questions (**BERDS**), where each example consists of a question and diverse perspectives associated with the question, sourced from survey questions and debate websites. On this data, retrievers paired with a corpus are evaluated to surface a document set that contains diverse perspectives. Our framing diverges from most retrieval tasks in that document relevancy cannot be decided by simple string matches to references. Instead, we build a language model based automatic evaluator that decides whether each retrieved document contains a perspective. This allows us to evaluate the performance of three different types of corpus (Wikipedia, web snapshot, and corpus constructed on the fly with retrieved pages from the search engine) paired with retrievers. Retrieving diverse documents remains challenging, with the outputs from existing retrievers covering all perspectives on only 33.74% of the examples. We further study the impact of query expansion and diversity-focused reranking approaches and analyze retriever sycophancy.

1 Introduction

Given a complex and contentious question (Xu et al., 2024), such as "Will ChatGPT do more harm than good?", a retrieval system should be able to surface diverse opinions in their top retrieval outputs. However, existing information retrieval (IR) tasks and systems mostly optimize for relevance to the question, ignoring diversity. Surfacing diverse documents can be useful to the users directly, but also can be improve retrieval-augmented language models (RALMs). Prompting large language models (LLMs) to generate an answer that encompasses diverse perspectives on its own is challenging (Sorensen et al., 2024; Hayati et al., 2023a), and retrieval-augmentation (Divekar and Durrett, 2024) can facilitate LLMs to generate more comprehensive answers that represent diverse perspectives.

Therefore, we study retrieval diversity, evaluating whether a set of top-scored outputs from the retrieval system cover diverse *perspectives* for an input query. We loosely define *perspective* as a particular attitude or viewpoint towards the question expressed towards the input query. Figure 1 shows examples of possible perspectives for different input questions. To tackle this task, we curate a **Be**nchmark for **R**etrieval **D**iversity for **S**ubjective questions, **BERDS**, which contains 3K complex questions, each paired with on average 2.3 perspectives. The questions come from three sources: survey question collection (Santurkar et al., 2023), debate question corpus (Wachsmuth et al., 2018), and newly scrapped data from a debate platform (Kialo) with argument maps.

Existing retrieval benchmarks are either evaluated through string match to reference answers (Karpukhin et al., 2020b) or exact match to the annotated document index in the corpus (Thakur et al., 2021; Xiao et al., 2024). We do not assume a corpus with gold document annotated and evaluate retrievers in an open-world setting (i.e., without assuming specific knowledge sources). To evaluate in this more realistic scenario, we develop an LLM-based automatic evaluator that determines if a document contains a certain perspective. We evaluate the performance of a retriever paired with a corpus. To quantify retrieval perspective diversity, we modify the evaluation metric used in multi-answer retrieval (Min et al., 2021b), where each question is paired with multiple valid factoid answers. MRECALL @ k measures whether a set of top-k documents cover all valid perspectives. Figure 1 provides example instances in **BERDS** and illustrates the proposed task.

Having constructed an evaluation framework, we test retrievers (Robertson et al., 2009; Karpukhin

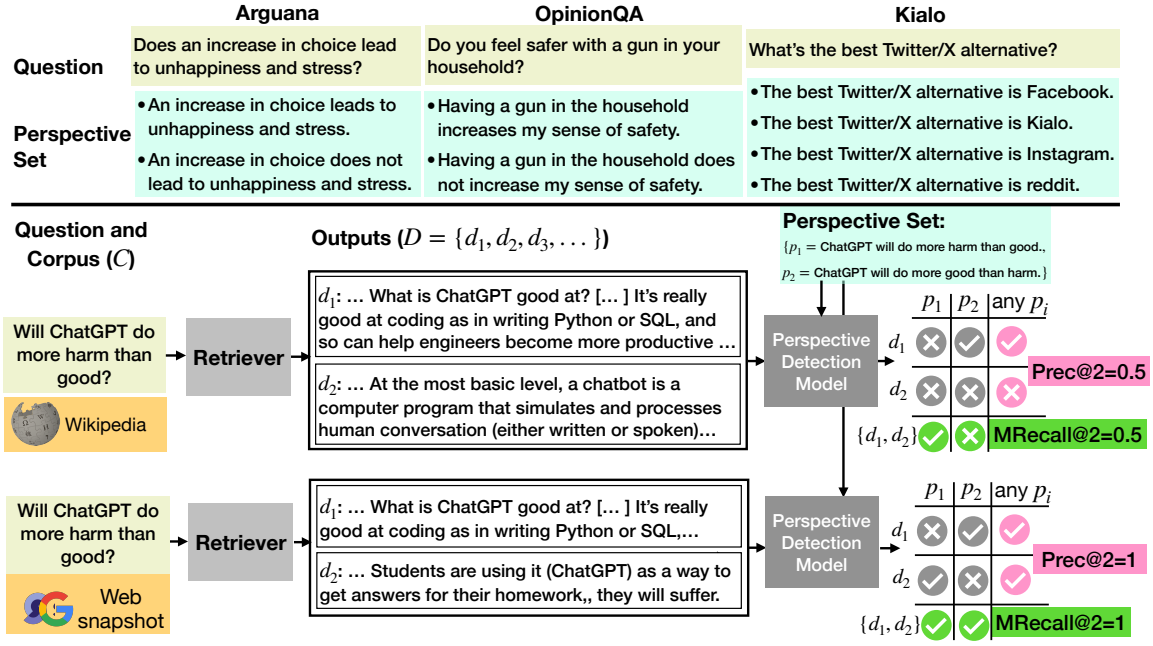

Figure 1: Overview of our benchmark and task. The upper part shows example instances. Each instance consists of a question and a perspective set. The lower part illustrates the task. The retriever will return a set of documents, and we evaluate whether the retrieved set (in this example, consisting of two documents) contains multiple answers or perspectives. MRecall measures the coverage of answers and perspectives, and precision measures whether each document contains any perspectives. A separate module (denoted as "Perspective Detection") is developed to determine whether a document contains a certain perspective. We experiment with retrieving from Web corpora instead of Wikipedia, which provides insufficient information for the task.

et al., 2020a; Izacard et al., 2022) and different corpus. Given the subjectivity of the questions we consider, Wikipedia alone is insufficient for answering many questions in our proposed datasets. Therefore, we use Sphere (Piktus et al., 2021), a subset of a web snapshot (CCNet (Wenzek et al., 2020)), and a retrieval corpus built on the fly for each question from the Google search engine output. We find that pairing a dense retriever with web corpora (Sphere and Google Search) yields more diverse outputs compared to other settings. Yet, our experimental result suggests that current retrievers surface relevant documents but cannot present document sets with diverse perspectives, even when retrieving from the web.

To enhance the diversity of the retrieval results, we implement simple re-ranking (Carbonell and Goldstein, 1998) and query expansion approaches (Mao et al., 2021). The former recomputes the scores for each document by penalizing documents that are similar to the documents that were retrieved before. The latter first generates multiple perspectives regarding the question using LLM (GPT-4) and uses them to guide diverse retrieval. Encouragingly, both methods showed strong gains on dense base retriever (Izacard et al., 2022). We further provide rich analysis studying the coverage of each corpus, retriever sycophancy, and whether retrievers prefer supporting or opposing perspectives to the input query.

We will release all our code, data, and trained model checkpoints to promote future work in studying retrieval diversity for complex questions.

2 Related Work

Diverse Perspectives Human society possesses pluralistic values (Sorensen et al., 2024) and exhibits diverse perspectives in various scenarios, including responses to survey questions (Santurkar et al., 2023; Durmus et al., 2023), news editorials (Liu et al., 2021), and stance with respect to debate or controversial questions (Chen et al., 2019, 2022b; Wan et al., 2024). Hayati et al. (2023b) looks into eliciting diverse perspectives from LLM on social norm topics. Wan et al. (2024) studies LLM generation when in-context documents provide contradicting statements. Concurrent work (Xu et al., 2024) studies evaluating perspective diversity of debatable questions. One way to encourage diversity in LLM output is retrieval augmentation, by providing documents with diverse perspectives. We focus on diverse retrieval

instead of generation from LLMs.

Retrieval Diversity Previous works have investigated retrieval diversity in terms of entity ambiguity (Clarke et al., 2008; Agrawal et al., 2009; Wang et al., 2017). Closer to our definition of diversity, Chen et al. (2019) propose a task *substantiated perspective discovery*, where systems are supposed to discover diverse perspectives taking stances against a given claim. Chen et al. (2022b) further argues for the need of a multi-perspective search engine that provides more direct and diverse responses. Despite these efforts, there is no existing benchmark for retrieving diverse perspectives in realistic corpora. To facilitate progress towards retrieving diverse perspectives for subjective questions, we formalize the task and construct a benchmark for evaluation, paired with automatic evaluation metrics. Concurrent work (Ziems et al., 2024) studies the bias of retrievers along a single polar axis (e.g. capitalism vs. socialism) for debatable questions and develops an unsupervised automatic metric.

Concurrent work (Zhao et al., 2024) explores the perspective-awareness of retrieval systems, where they repurpose existing datasets to build a retrieval benchmark where queries are supplemented with a perspective. While sharing a similar goal, they evaluate retrieval with perspectives as input and measure if systems follow them, and in our settings the retrievers are not given target perspectives. Furthermore, our work expands knowledge sources significantly using a web corpus, while their work focuses on much smaller corpora.

Stance Detection (Mohammad et al., 2016; Al-Dayel and Magdy, 2021) evaluates if a piece of text supports a target (an entity, idea, opinion, claim, etc.). Sen et al. (2018) investigate multi-perspective questions in health information and propose methods to classify search results according to their stance towards the question. Wan et al. (2024) determines the stance of search results to contentious questions using LLMs. Our perspective detection (Section 3) subsumes stance detection, which classifies the target as positive or negative (binary). Perspective detection can be generalized to any number or type of perspectives.

Retrieval from the Web Piktus et al. (2021) proposes to retrieve from the web to account for the open-domain nature of real-world knowledge-intensive applications. Previous works explored retrieval with search engines (e.g. Google Search API) to help fact-checking (Chen et al., 2024), open-domain question-answering (Lazaridou et al., 2022), answer-editing (Gao et al., 2023a), and retrieval-augmented generation (Li et al., 2023). Other works have investigated attributing LM outputs to evidence documents retrieved from the web (Malaviya et al., 2023; Gao et al., 2023b).

Dataset	# Total (Dev./Test)	# p	$\lvert p \rvert$
Arguana	1000 (250/750)	2.0	12.71
OpinionQA	1176 (294/882)	2.0	16.51
Kialo	1032 (258/774)	2.9	12.49
All	3208 (802 / 2407)	2.28	13.84

Table 1: Dataset statistics. The second column denotes the number of examples. We spare 25% of each dataset as the development (Dev.) set, and use the remainder as the test set. # p is the average number of perspectives per question, and $|p|$ denotes the average number of words (by NLTK `word_tokenize`) in each perspective.

3 Task Formulation

The retrieval system will be provided a question q and a corpus C, and return a ranked list of k documents, $D = \{d_1, d_2, ..., d_k\}$, where $d_i \in C$ and d_i is the i-th most relevant document in the corpus to the question q.

Background: Multi-answer Retrieval In earlier work (Robertson et al., 2009; Voorhees et al., 1999; Chen et al., 2017; Lee et al., 2019), each question q is assumed to have a single gold answer a. Thus, the system is evaluated by assigning a binary label, RECALL @ k, indicating if any document d_i in the retrieved document set D contains a.

Min et al. (2021b) extended retrieval task to cover questions that have multiple valid answers. Here, each question q is paired with gold answer set $\{a_1, a_2, ...a_m\}$ where each gold answer a_i is a short string. The goal is to retrieve a diverse set of documents such that the retrieval output will contain all m answers. New evaluation metric MRECALL @ k is defined as (1) when number of answers $m > k$, MRECALL @ k = 1 if the retrieved document set D contains all k answers. (2) when $m \leq k$, MRECALL @ k = 1 if the retrieved document set D contains all m answers. Otherwise MRECALL @ k = 0. Simple string matching determines if a document d_i contains answer a.

Ours: Multi-perspective Retrieval We consider questions on subjective topics that could be approached from different perspectives, and there is no fixed or complete set of reference answers. We

assume a set of m perspectives $\{p_1, ..., p_m\}$ for addressing question q. Retrievers should output a diverse set of documents covering all m perspectives.

Evaluation Challenges Deciding if a document contains an answer can be approximated by simple string matching, but deciding if a document supports a perspective is nontrivial. The most accurate metric would be human judgments. However, this is not scalable, so we build a model approximating human judgment (presented in Appendix A). We refer to this task of determining if a document contains a certain perspective as "Perspective Detection". We do not use off-the-shelf stance detection models (Zhang et al., 2023a; Alturayeif et al., 2023) as the perspectives might not be binary. A single document could contain more than one perspective.

Evaluation Metrics We use MRECALL @ k from prior work (Min et al., 2021b). The key difference is instead of checking whether a document contains any of answer $\{a_1, a_2, ...a_m\}$, we check whether a document supports any of perspectives $\{p_1, p_2..., p_m\}$. We also report PRECISION @ k, the percentage of documents in the retrieved document set D that contain any perspectives. This measures the relevance of the retrieved documents to the question.

This evaluation setup does not assume annotated "gold" documents or answers for each question, enabling comparing different corpora. We discuss the effects of corpora in Section 7.

4 The BERDS Benchmark

The BERDS Benchmark aims to measure retrieval diversity for questions that are opinionated or invite diverse perspectives. A single instance in BERDS will consist of a question q and a set of m perspectives $\{p_1, p_2..., p_m\}$. Each perspective should be relevant to the question and distinct. We provide the data selection and construction process below.

4.1 Data Construction

We aim to look for questions that elicit diverse valid perspectives naturally. Questions related to debate topics would be good candidates since they would induce at least two contradicting, valid viewpoints. Another natural choice is survey questions, where people would express differing opinions. We thus collect questions from a debate website (`kialo.com`) and a dataset of survey questions, OpinionQA (Santurkar et al., 2023). We also generate questions from Arguana (Wachsmuth et al., 2018), a dataset of contradicting arguments. We provide examples for each dataset in BERDS in Figure 1, and data statistics in Table 1. For both Arguana and OpinionQA, we only consider two contradictory perspectives, one supporting and one opposing. These perspectives may not be exhaustive but cover a range of arguments for the question. We additionally provide examples of the raw data and generated data in Table 10.

Corpus	Source	# pas.	# doc.
Wikipedia	Wikipedia	22 M	5.9 M
Sphere	CCNet	906 M	134 M
Google Search output		4461.7	74.2

Table 2: Retrieval corpus statistics. # pas. is the number of passages; # doc. is the number of documents, where each document is a 100-word passage. For Google Search output, we construct a corpus for each example, and we report the average statistics across examples.

Arguana is a dataset on the task of retrieving the best counterargument. As each instance in the Arguana dataset contains contradicting arguments (the query and its corresponding counterargument), it naturally induces two distinct perspectives. We generate questions from the contradicting arguments, and then generate corresponding perspectives for each argument, using GPT-4.

Kialo We collect 1,032 questions from a debate website (`kialo.com`), which mostly consist of Yes/No questions (91%). For Yes/No questions, the website provides supporting statements converted from the questions. We consider the provided statement as one perspective and generate from it an opposing perspective using GPT-4. For the rest of the questions (9%) that naturally invite more than two perspectives (e.g. What's the best Twitter/X alternative?), we generate corresponding perspectives (e.g. The best Twitter/X alternative is Kialo) for each answer (e.g. Kialo) in the website.

OpinionQA is a dataset of survey questions on various topics including privacy, political views, and health, targeted toward the citizens of the United States. The dataset originates from the annual Pew American Trends Panel survey. The dataset contains a wide range of survey questions, such as those asking about personal experiences (*Regardless of whether or not you own a gun, have you ever fired a gun?*). We filter such questions by prompting GPT-4, removing 22.7% of the ques-

tions (345 out of 1521 in total). The survey questions come with multiple options, with each representing a certain degree of support for the question. Conditioned on the survey question and the options, we convert each survey question into a natural language question and generate one supporting and one opposing perspective with GPT-4.

We use `gpt-4-0613` for all the data generation. All the input prompts for generation can be found in Appendix B.1. The total cost for generating the data is about $100 USD.

4.2 Retrieval Corpus

We consider three types of retrieval corpus: Wikipedia, web snapshot, and the web itself. While prior work on factoid QA (Chen et al., 2017; Lee et al., 2019; Karpukhin et al., 2020a) focused on retrieving from Wikipedia only, our task, designed to uncover a wide range of opinions and perspectives, will benefit from retrieving from a web corpus.

Wikipedia We consider the Wikipedia dump processed by (Karpukhin et al., 2020a), consisting of disjoint 100-word segments.

Web Snapshot: Sphere Prior works on retrieval usually adopt Wikipedia as the main retrieval corpus. However, Wikipedia alone cannot cover all the information required (Redi et al., 2020). This is even more likely as our questions are inherently opinionated and Wikipedia texts are supposed to be neutral. We propose to use the Sphere (Piktus et al., 2021) corpus, which is a subset of CCNet (Wenzek et al., 2020), as the knowledge source. One advantage of retrieving from web corpus over using a commercial search engine (e.g. Google Search) results is reproducibility.

Online Web Search: Google Search API Instead of building a static corpus and retrieving documents from it, one can use a web search engine to retrieve new documents on the fly. This enables the system to access bigger, more up-to-date source documents. The downside of using web search API (Nakano et al., 2021; Yoran et al., 2023) is the lack of reproducibility as well as potentially returning documents from less credible sources.[1]

We implement web search by retrieving documents from Google Search API.[2] We first obtain the raw HTML of top hundred documents, which are then processed and converted to plain text using html2text[3] and readability,[4] following Chen et al. (2022a, 2024). We split the processed documents into 100-word segments and take all the segments as the corpus. Unlike Wikipedia and Sphere, we build a separate corpus for each question.

5 Retrieval Approaches

We consider a suite of retrievers (sparse and dense) on our dataset. As these retrievers do not encourage the diversity of their output, we implement a reranking method and a query expansion method.

Base Models We implement BM25 (Robertson et al., 2009), a widely adopted bag-of-word retrieval function considering token-matching between questions and documents. We use the LuceneSearcher of Pyserini package to build the Wikipedia corpus. For the Sphere corpus, we use the faiss index provided by Piktus et al. (2021).

We consider four dense retrievers, DPR (Karpukhin et al., 2020a), CONTRIEVER (Izacard et al., 2022), textscTart (Asai et al., 2023) and NV-Embed-v2 (Lee et al., 2024). DPR is a dual encoder trained with contrastive learning on the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019). CONTRIEVER is an encoder trained with a self-supervised contrastive objective. DPR uses separate encoders for document and question embeddings, and CONTRIEVER uses a shared encoder. Both models are initialized from a BERT-base encoder. We use the version of CONTRIEVER further fine-tuned on the MSMARCO dataset (Nguyen et al., 2016) after self-supervised learning, which performed better in the original paper. TART (Asai et al., 2023) is trained on a large-scale retrieval dataset with instructions. We implement TART-full, the cross-encoder version, and the top hundred documents from CONTRIEVER are used as initial candidates. NV-Embed-v2 is the state-of-the-art text embedding model.[5]

[1] Prior work (Chen et al., 2024) studied the reproducibility of search API results, comparing URLs between the sets of documents retrieved at different timestamps with the same query. They report only 30% of URLs overlap when queried two months apart.

[2] https://serper.dev/

[3] https://pypi.org/project/html2text/

[4] https://github.com/buriy/python-readability

[5] Ranked first in the MTEB (Muennighoff et al., 2022) benchmark as of Oct 15, 2024. We do not include results on Sphere due to constraints on compute. Typical dual encoder models like DPR use a BERT-base encoder (110M parameters) with an embedding size of 768, while NV-Embed-v2 has 7B parameters and an embedding size of 4096, significantly increasing the cost of generating document indices.

Corpus	Model	Datasets						Macro-Average		
		Arguana		Kialo		OpinionQA				
		Prec.	MRec.	Prec.	MRec.	Prec.	MRec.	Prec.	MRec.	Prec.@ 1
Wiki	BM25	22.27	10.8	23.82	10.85	10.39	2.95	18.83	8.20	21.72
	DPR	18.88	7.73	19.48	7.62	13.63	3.29	17.33	6.21	20.22
	CONTRIEVER	35.79	19.2	32.97	15.63	23.99	8.28	30.92	14.37	36.85
	TART	37.84	21.07	33.97	16.80	25.36	8.40	32.39	15.42	38.62
	NV-Embed-v2	40.69	23.20	38.99	22.27	28.80	8.13	36.16	17.87	43.15
Sphere	BM25	44.96	27.6	43.33	24.16	50.11	22.45	46.13	24.74	51.26
	DPR	10.80	4.00	13.85	5.43	8.34	2.15	11.00	3.86	11.65
	CONTRIEVER	58.21	38.27	54.96	28.94	52.77	24.72	55.31	30.64	58.67
	TART	**65.47**	**43.60**	56.47	32.80	**63.25**	**31.60**	**61.73**	**36.00**	65.16
Google Search Output	BM25	35.15	21.07	35.43	22.61	41.27	21.09	37.28	21.59	43.02
	CONTRIEVER	52.32	27.6	50.31	28.94	48.91	22.79	50.51	26.44	57.56
	TART	53.33	30.13	50.48	27.87	50.83	24.67	51.55	27.56	59.07
	NV-Embed-v2	62.40	39.87	**64.19**	**37.60**	56.88	28.67	61.16	35.38	**67.87**

Table 3: Performance of retrievers on the test split of BERDS.

Re-ranking We apply the Maximal Marginal Relevance (MMR) method (Carbonell and Goldstein, 1998), which uses the following selection criteria for re-ranking:

$$\arg\max_{D_i \in R \setminus S} [\lambda Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j)] \quad (1)$$

where R denotes a set of top retrieved documents considered for re-ranking ($|R| = 100$ in our case), S denotes the document set that is already selected, and λ is a hyperparameter, which we tune on the development split.[6] Sim_1 is the score from the retriever. Sim_2 is the cosine similarity between two document embeddings computed using the Universal AnglE Embedding (Li and Li, 2023). We choose this model as it is the top-ranked open model on the MTEB leaderboard (Muennighoff et al., 2023) with less than 1B parameters (335M). Compact models are preferable as we compute document embeddings for the top hundred documents per example. The experimental details can be found in Appendix B.6.

Query Expansion Prior work on factoid QA (Mao et al., 2021) showed that first generating an answer with LLM and then using the generated answer to augment the question can improve retrieval. Similarly, we first generate multiple (n) perspectives on the given question using LLMs (in our case `gpt-4-0613`), and query the retriever with each generated perspective.[7] We take the top document from each document set retrieved using the generated perspectives in a round-robin fashion k times to form a list of $n * k$ documents. We then take the top k unique documents from this list, skipping duplicates, as the same documents could be obtained from different queries. We follow the order in which the perspectives are generated. We include a visual illustration in Appendix B.3. The prompts for generating the perspectives and examples are in Appendix B.3 and Appendix B.7.

6 Experimental Results

We report MRECALL @ 5 and PRECISION @ 5,[8] as defined in Section 3. All evaluation metrics require judging whether each retrieved document contains a perspective (defined as perspective detection, and more details can be found in Appendix A). We use a fine-tuned Mistral-7B model as the evaluator.[9]

Table 3 presents the performance of base retrievers on BERDS using three retrieval corpus described in Section 4.2. We do not report DPR results on Google Search outputs as we find it does not generalize well beyond the Wikipedia corpus.

Unlike prior work (Karpukhin et al., 2020a) which assumes a gold corpus, our retrieval corpus might not contain documents representing diverse perspectives on the provided question. Thus, the performance metrics reflect both the coverage of the corpus and the retriever's ability to surface documents supporting diverse perspectives. We analyze the coverage of each corpus in Section 7.

Effect of Corpus. Comparing retrievers operating on Wikipedia and Sphere, most retrievers achieve higher MRECALL and PRECISION when

[6] We normalize the Sim_1 scores to be between 0 and 1, and consider λ values of [0.5, 0.75, 0.9, 0.95, 0.99].

[7] n is not predetermined; GPT-4 could generate varying numbers of perspectives. n = 6.1 on average.

[8] We additionally report the results for k = 10 in Appendix C.1, where very similar trends hold.

[9] We will open-source this model to promote future work.

Corpus	Retriever	Default		+ Re-ranking		+ Query Expansion	
		Prec.	MRec.	Prec.	MRec.	Prec.	MRec.
Wiki	BM25	18.83	8.20	18.66 (-0.9%)	8.03 (-2.1%)	**21.31** (+13.2%)	**10.39** (+26.7%)
	DPR	17.33	6.21	16.62 (-4.1%)	6.06 (-2.5%)	**17.78** (+2.6%)	**6.65** (+7.1%)
	CONTRIEVER	30.92	14.37	27.89 (-9.8%)	**15.54** (+8.2%)	**33.23** (+7.5%)	14.84 (+3.3%)
Sphere	BM25	**46.13**	24.74	45.43 (-1.5%)	25.38 (+2.6%)	44.16 (-4.3%)	**26.24** (+6.1%)
	DPR	**11.00**	**3.86**	8.16 (-25.8%)	2.60 (-32.6%)	6.20 (-43.6%)	2.01 (-48.0%)
	CONTRIEVER	**55.31**	30.64	54.39 (-1.7%)	33.55 (+9.5%)	54.91 (-0.7%)	**33.74** (+10.1%)
Google Search Output	BM25	37.28	21.59	36.75 (-1.4%)	21.18 (-1.9%)	**44.87** (+20.3%)	**24.96** (+15.6%)
	CONTRIEVER	50.51	26.44	49.33 (-2.3%)	**26.56** (+0.4%)	**52.60** (+4.1%)	24.83 (-6.1%)

Table 4: Performances on BERDS after re-ranking and query expansion. Each cell reports the macro-average over three datasets in BERDS. **Bolded** numbers are the best performance in each row. Blue numbers indicate increases compared to the default setting.

retrieving from Sphere, except DPR. We hypothesize that DPR is trained only on the documents from the Wikipedia corpus so that it is not generalizing to the web corpus. Taking web search results (Google Search) as a corpus shows improved diversity over Wikipedia, but lags behind retrieving from the entire web (Sphere). The gap in the corpus size might cause this – we only take the top 100 Google Search results and build corpora with them (resulting in a corpus size of 4.4K passages on average), while Sphere contains 906M passages.

Comparing Retrievers. Across all experimental settings, TART achieves the highest diversity, followed by CONTRIEVER, BM25 and DPR. BM25 generalizes better than DPR, which is fine-tuned on NQ with Wikipedia corpus. TART achieves higher diversity than CONTRIEVER, suggesting that specifying the task with instructions helps diversity. Even the best setting (TART on Sphere) only achieves an average MRECALL of 32.8. This indicates even the best retriever struggles to retrieve a comprehensive set of documents. We investigate the number of documents needed to be retrieved to cover diverse perspectives in Appendix C.2. PRECISION @ 1 is as high as 58.67, showing that retrievers could surface one of the perspectives.

Results: Approaches to Improve Diversity We report the results of re-ranking and query expansion approaches in Table 4. We report the results with three retrievers (BM25, DPR, Contriever), excluding TART and NV-Embed2 for computational costs. Re-ranking improves diversity on four of eight retriever-corpus settings, at the cost of PRECISION for all settings. The limited performance gains could be because semantic dissimilarity does not always infer differing perspectives. Query expansion improves the PRECISION for five settings and boosts MRECALL for six settings (over 10% for three of them). This indicates that GPT-4 could diversify the search space by providing queries hinting different perspectives to the retrievers. Future work can explore combining both approaches.

7 Analysis

Do retrievers prefer documents with supporting or opposing perspectives to the question? Retrievers often fail to surface documents that cover comprehensive perspectives. Can we characterize the perspective that is more likely to be surfaced? We focus on questions with only two perspectives, one taking the supporting stance towards the question and the other opposing.[10] We report whether the top five documents obtained by the retrievers cover only supporting, opposing, both, or neither perspectives in Figure 2. We observe that all retrievers retrieve the supporting perspectives significantly more often when they fail to retrieve both.

Retriever Sycophancy. Sycophancy refers to one's tendency to tailor their response to please their interaction partner. Prior work has explored how LLMs exhibit sycophancy (Perez et al., 2023). We examine if similar behavior is observed with the retrievers; i.e. do retrievers favor documents that share the perspective with the question? For example, would the retrievers surface more documents supporting a perspective, "ChatGPT will do more harm than good.", if the perspective is used as the question for the retrievers, compared to using "Chat GPT will do more good than harm" as the question? We conduct a controlled study where we input either the supporting or opposing perspectives as the query to the retriever, and observe the

[10] We identify this by prompting GPT-4; details and prompts can be found in Appendix B.8.

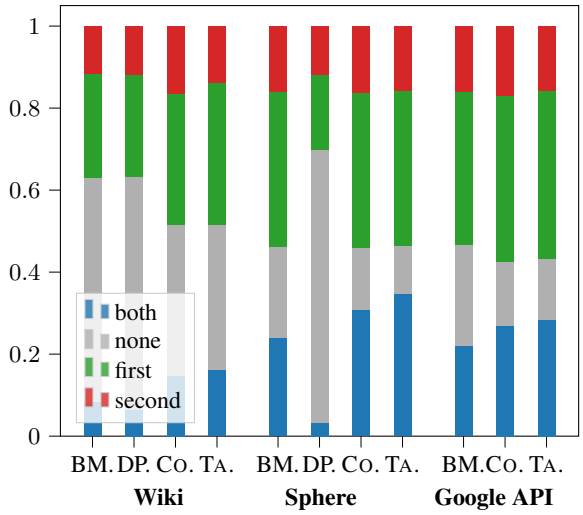

Figure 2: We report whether retrievers prefer the supporting or the opposing perspectives. Results are computed on the top-5 documents aggregated over three datasets. All retrievers favor the supporting perspectives. BM. denotes BM25 retriever, DP. denotes DPR, and CO. denotes CONTRIEVER. TA. denotes TART and Wiki. denotes the Wikipedia corpus.

changes to their leaning.

We quantify leaning towards supporting perspectives to be $\Delta = (p - n)/p$, where p and n are the percentages of documents containing supporting and opposing perspectives, respectively. Greater Δ indicates the retrievers favor the supporting perspectives more. We present the results on top 5 documents retrieved from the Wikipedia corpus in Figure 3.[11] Querying the retrievers with supporting perspectives increases Δ across the board, and retrieving with opposing perspectives decreases Δ. Retrievers tend to favor perspectives that they are prompted with. This partially explains the gains from the query expansion.

Is retrieval diversity limited by the corpora? Unlike prior work, we do not assume a gold corpus that contains target information. The best MRECALL @ 5 on BERDS achieved by any approach on average is below 35. How much of the failure is caused by the limitation of the corpus or that of retrievers? We answer this question by computing MRECALL over a larger document set, where we combine the top 100 retrieved documents from all retrievers for each corpus. If this number nears 100, we could theoretically find an optimal subset of k documents for each question that achieve MRECALL @ k close to 1. We present the results in Table 5. Only the Wikipedia corpus contain insufficiently diverse information, achieving MRECALL of 62% when combining three retriever outputs. In contrast, the web corpora do not limit retrievers' performances in diversity. This result hints at great room for improvement for the retrievers.

[11] Results on the other corpora show the same trends, except for DPR on the web corpus.

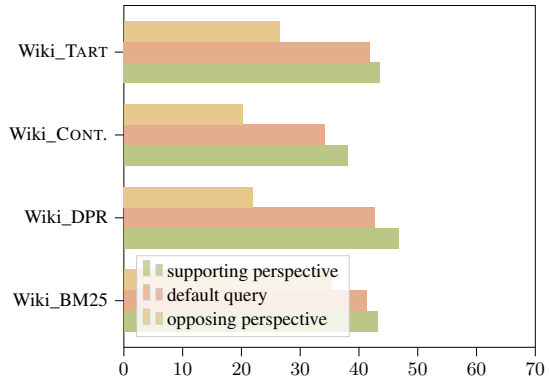

Figure 3: Retriever sycophancy. X-axis reports the propensity of retrieval output to prefer positive, the $\Delta = \frac{p-n}{p}$, where p is the number of retrieved documents that contain supporting perspective and n is the number of retrieved documents that contain opposing perspective. We label each setting as `[corpus]_[retriever]`. In all retrievers, using supporting perspective as query leads to higher supportive leaning, and using opposing perspective as query leads to lower supportive leaning.

Corpus	Arguana	Kialo	OpinionQA	Average
Wiki.	71.87	61.24	55.10	62.74
Sphere	92.80	85.27	89.46	89.18
Google	82.67	82.69	77.66	81.01

Table 5: MRECALL over the union of the top 100 retrieved documents from all retrievers for each corpus. These values are substantially higher than MRECALL @ 5, suggesting the corpora (especially Sphere) have good information coverage.

8 Conclusion

We study retrieval diversity for complex, subjective questions. We curate a benchmark BERDS that evaluates retrieval diversity, provides automatic metrics, and establishes the performances of retriever baselines. Our experiments show that current retrievers cannot retrieve comprehensive information in the proposed setting.

After retrieving diverse documents, summarizing such rich information into a coherent response (Laban et al., 2022; Huang et al., 2023; Zhang et al., 2023b) could be useful for the users.

Limitations

- The perspective set in our dataset is mostly binary (the average number of perspectives is 2.9 for Kialo and 2.0 for Arguana and OpinionQA). More fine-grained perspectives could be explored.
- Our dataset covers debate and opinion survey questions. Exploring other domains like healthcare would be an interesting future direction.
- The evaluator LM has 7B parameters, making evaluation slow. Evaluating one dataset in BERDS using the evaluator takes about three hours on an NVIDIA A40 GPU. A vLLM (Kwon et al., 2023) implementation reduces the evaluation time to 15 minutes. We do not discuss the efficiency aspect as we focus on the accuracy of the evaluator. Future work could build more time-efficient evaluators that are as accurate.
- The majority of data is GPT-4 generated. We do not have a guarantee of the data quality, but manually inspecting a small portion of the data indicates that it is not a serious issue.

References

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the second ACM international conference on web search and data mining*, pages 5–14.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675, Toronto, Canada. Association for Computational Linguistics.

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.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. 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 3569–3587, Mexico City, Mexico. Association for Computational Linguistics.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022a. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Sihao Chen, Siyi Liu, Xander Uyttendaele, Yi Zhang, William Bruno, and Dan Roth. 2022b. Design challenges for a multi-perspective search engine. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 293–303, Seattle, United States. Association for Computational Linguistics.

Charles LA Clarke, Maheedhar 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.

Abhishek Divekar and Greg Durrett. 2024. Synthesizrr: Generating diverse datasets with retrieval augmentation. *Preprint*, arXiv:2405.10040.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and

Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023a. How far can we extract diverse perspectives from large language models? criteria-based diversity prompting! *arXiv preprint arXiv:2311.09799*.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023b. How far can we extract diverse perspectives from large language models? criteria-based diversity prompting! *ArXiv*, abs/2311.09799.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kung-Hsiang Huang, Philippe Laban, Alexander R Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. *arXiv preprint arXiv:2309.09369*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020a. 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.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020b. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Xiang Chen, and Caiming Xiong. 2022. Discord questions: A computational approach to diversity analysis in news coverage. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5180–5194, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jingyuan Wang, Jian-Yun Nie, and Ji-Rong Wen. 2023. The web can be your oyster for improving language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 728–746, Toronto, Canada. Association for Computational Linguistics.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Siyi Liu, Sihao Chen, Xander Uyttendaele, and Dan Roth. 2021. MultiOpEd: A corpus of multi-perspective news editorials. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4345–4361, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. *arXiv preprint arXiv:2309.07852*.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online. Association for Computational Linguistics.

Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, et al. 2021a. Neurips 2020 efficientqa competition: Systems, analyses and lessons learned. In *NeurIPS 2020 Competition and Demonstration Track*, pages 86–111. PMLR.

Sewon Min, Kenton Lee, Ming-Wei Chang, Kristina Toutanova, and Hannaneh Hajishirzi. 2021b. Joint passage ranking for diverse multi-answer retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6997–7008, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

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*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard, Patrick Lewis, Barlas Oğuz, Edouard Grave, Wen-tau Yih, et al. 2021. The web is your oyster-knowledge-intensive nlp against a very large web corpus. *arXiv preprint arXiv:2112.09924*.

Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. 2020. A taxonomy of knowledge gaps for wikimedia projects (second draft). *arXiv preprint arXiv:2008.12314*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

Anirban Sen, Manjira Sinha, Sandya Mannarswamy, and Shourya Roy. 2018. Stance classification of multi-perspective consumer health information. In *Proceedings of the ACM India joint international conference on data science and management of data*, pages 273–281.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv*, abs/2104.08663.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? *arXiv preprint arXiv:2402.11782*.

Xiaojie Wang, Ji-Rong Wen, Zhicheng Dou, Tetsuya Sakai, and Rui Zhang. 2017. Search result diversity evaluation based on intent hierarchies. *IEEE Transactions on Knowledge and Data Engineering*, 30(1):156–169.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. 2024. Rar-b: Reasoning as retrieval benchmark. *Preprint*, arXiv:2404.06347.

Rongwu Xu, Xuan Qi, Zehan Qi, Wei Xu, and Zhijiang Guo. 2024. Debateqa: Evaluating question answering on debatable knowledge. *arXiv preprint arXiv:2408.01419*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *ArXiv*, abs/2310.01558.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Yangyang Li, and Liwen Jing. 2023a. Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*.

Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, et al. 2023b. Fair abstractive summarization of diverse perspectives. *arXiv preprint arXiv:2311.07884*.

Xinran Zhao, Tong Chen, Sihao Chen, Hongming Zhang, and Tongshuang Wu. 2024. Beyond relevance: Evaluate and improve retrievers on perspective awareness. *arXiv preprint arXiv:2405.02714*.

Caleb Ziems, William Held, Jane Dwivedi-Yu, and Diyi Yang. 2024. Measuring and addressing indexical bias in information retrieval. *arXiv preprint arXiv:2406.04298*.

A Identifying whether a perspective is represented in a document

Our primary goal is developing and evaluating retrieval methods for uncovering diverse perspectives. To this end, we need a system that can identify whether the retrieved document set contains a perspective. Unlike simple factoid QA setting, where we can evaluate whether a document contains an answer by string match (Lee et al., 2019; Min et al., 2021a,b), evaluating whether a document contains a perspective is nontrivial, a research question on its own (Sen et al., 2018; AlDayel and Magdy, 2021). In this section, we aim to build an efficient model for this subtask, which is a necessary component for the evaluation of diverse retrieval models.

Task and Metric Given a perspective and a document, the model assigns a binary label y indicating whether the document contains the perspective. Then, we measure the performance of a model by comparing its prediction with the reference label, measuring accuracy and F1 score.

Reference Annotation For each document-perspective pair, we consider human judgment as the gold label.

The human-labeled test set contains 542 document-perspective pairs, which are annotated by the authors of this paper. We compute inter-annotator agreement on 176 examples, yielding a Cohen Kappa agreement of 0.56. 16.8% of the data receive a positive label. The examples are taken from top five retrieval results from the Sphere corpus with BM25 (Robertson et al., 2009) and DPR (Karpukhin et al., 2020a) retriever described in Section 5, and equally split between three datasets. Details on the construction of the human reference set can be found in Appendix B.4.

A.1 Evaluated Approaches

Random We report a baseline that randomly assign predictions $y = 1$ or 0 according to the label distribution of each dataset.

Off-the-shelf NLI models We could interpret "a document d_i contains perspective p_j" as "a premise d_i supports a hypothesis p_j". Then it becomes natural to use NLI models to identify whether a perspective is represented in a document. We use a T5-11B model provided by Honovich et al. (2022).

Prompted LMs Instead of using an NLI model, we can prompt large-scale LLM to decide whether a document contains a perspective. From the open-source LMs, we choose Llama-2-chat (Touvron et al., 2023), an instruct fine-tuned version of Mistral-7B (Jiang et al., 2023), Zephyr-7B (Tunstall et al., 2023), and Gemma-7B (Team et al., 2024). Lastly, we use GPT-4 (`gpt-4-0613`). To reach a better absolute performance, we additionally experiment with prompting with one in-context example for GPT-4.

Models	% Pos	Acc	F1
Reference human	16.8	-	-
Another human	22.7	85.2	65.8
Random	16.8	72.1	16.9
NLI model (T5-11B)	7.2	85.2	38.5
Llama-2-13b-chat (zero-shot)	31.1	70.7	42.6
Llama-2-70b-chat (zero-shot)	27.4	77.7	52.6
Zephyr (zero-shot)	22.4	80.6	53.3
Gemma (zero-shot)	58.4	48.9	37.2
Mistral-7B (zero-shot)	19.2	82.2	52.0
GPT-4 (zero-shot)	17.4	**87.6**	**65.6**
GPT-4 (one-shot)	15.4	**89.1**	**67.8**
Mistral-7B (fine-tuned)	15.8	87.6	62.2

Table 6: Perspective detection results on human-labeled data (n=542). "Another human" performance is computed by taking one set of human labels as the ground truth and the other as predictions on the 176 examples where we compute inter-annotator agreement.

Fine-tuned LMs Prompting GPT4 showed promising results in this task, but it would require additional API calls, which incur extra cost, whenever we would like to evaluate new systems on BERDS. We thus fine-tune the Mistral-7B model with GPT-4 predictions with one in-context example as labels. The fine-tuning dataset contains 3K examples corresponding to 50 unique questions, and implementation details can be found in Appendix B.5.

A.2 Results

We report the results on the human-labeled document-perspectives pairs in Table 6. The off-the-shelf entailment model (T5) does not work well, whereas GPT-4 is the best model. Fine-tuned Mistral-7B obtains the best results in open-source models, surpassing Llama-2-70b-chat. The prediction from fine-tuned Mistral-7B matches that of GPT-4 on 94% of instances.

As the fine-tuned Mistral-7B comes close to GPT-4 performance, we use this model as the automatic metric for evaluating retrieval diversity, providing an open-source model that enables reproducible, cheaper evaluation.

Corpus	BM25	DPR	CONTRIEVER
Wikipedia	0.99	0.95	0.75
Sphere	0.99	0.99	0.95
Google Search Output	0.95	-	0.95

Table 7: The λ value in Equation (1) for different retriever-corpus setting.

B Experimental Details

B.1 Prompts and Examples for Data Generation

Prompts for generating each dataset are shown in Tables 8, 9 and 11. Examples showing the generated data and data taken from the source can be found in Table 10.

B.2 Prompts for Perspective Detection

Prompts we use for perspective detection are shown in Table 14. We prompt all models with the prompt in the top row in Table 14, and additionally report results prompting GPT-4 with the in-context example shown in the bottom row (the last row and the best performance in Table 6). The labels of the data used to fine-tune Mistral-7B are predicted with the one-shot prompt.

B.3 Prompts for Query-Expansion

Prompts we use to generate perspectives for query-expansion are shown in Table 12.

B.4 Details on Constructing Perspective Detection Test Set

Each example in the test set consists of a perspective and a document. We randomly select 10 questions with two perspectives from each of the three datasets, yielding a total of 60 perspectives. For each question, we take the top five retrieved documents from either BM25 or DPR results obtained from Sphere. Thus there are ten documents to test for each perspective, resulting in 600 document-perspective pairs. We then check for near-duplicate documents and filter out 58 examples, leaving us with 542 examples.

B.5 Details of Fine-tuning Mistral-7B Models

We fine-tune the Mistral-7B model with LoRA (Hu et al., 2021) on eight NVIDIA A40 GPUs. We train the model on 3K examples of GPT-4 labeled data on perspective detection. The training dataset is constructed using the same process as described in Appendix B.4, except on 50 questions, yielding five times more data.

The model is trained using an AdamW (Loshchilov and Hutter, 2018) optimizer with a linear learning schedule. We perform a hyperparameter search on learning rates (1e-4 2e-4 1e-5 2e-5 3e-5 5e-5), warmup ratio (0.1 0.3 0.5), number of training epochs (0.5, 1, 2, 3, 4), and weight decay (1e-4, 0.001, 0.01, 0). We evaluate each checkpoint on a separate development set of 100 examples annotated by the author and select the model with the highest F1 score. The final hyperparameters we decided on are learning of 1e-4, a warmup ratio of 0.1, 2 epochs of training, and a weight decay of 0.001.

B.6 Detail of Re-ranking Experiments

We compute the Sim_2 scores in Equation (1) by taking the cosine similarity between two document embeddings computed by a sentence embedding model, the UAE-Large-V1 model on huggingface [12]. We follow their documentation for non-retrieval tasks and set the `pooling-strategy ='cls'`. We do not specify any prompt for the model so the embeddings are only on the documents themselves.

Sim_1 are obtained from the implemented retriever model, and thus they could be of different scales. We then normalize the Sim_1 scores to be between [0, 1] by dividing Sim_1 by the maximum score obtained on the whole dataset, and do a grid search on λ over values [0.5, 0.75, 0.9, 0.95, 0.99]. The actual values we choose for each retriever-corpus setting are shown in Table 7.

B.7 Examples of Generated Perspectives for Query Expansion

We present examples of perspectives generated by GPT-4 on each dataset in Table 13. Each perspective has a name separated by "_", and the corresponding text. The average length of perspectives is 43.98 words.

B.8 Details of Determining Supporting and Opposing Perspectives

We identify the supporting and opposing perspectives by prompting `gpt4-0613`, and the exact prompt can be found in Table 15. We only perform such a process for examples with two perspectives.

[12] https://huggingface.co/WhereIsAI/UAE-Large-V1

Setting	Prompt
Question Generation (2-shot)	Document 1: Natural habitats being are destroyed A tougher approach to the protection of animals is needed to prevent their natural habitats from being destroyed by locals. As humans expand their agricultural activity in Africa they are destroying the environments of endangered animals and pushing others towards being endangered. Due to an increase in large scale cotton plantations and food crops, the West African lion has seen a marked decrease in population; numbering less than 400 in early 2014 [1] . Tougher protection, such as fencing off areas from human activity, has been suggested and has seen success in South Africa [2] . [1] BBC, "Lions 'facing extinction in West Africa'" [2] Morelle,R. "Fencing off wild lions from humans 'could save them'" Document 2: Fewer human deaths Fewer large beasts will lead to fewer deaths in Africa. Some endangered animals are aggressive and will attack humans. Hippopotamuses kill in excess of three hundred humans a year in Africa, with other animals such as the elephant and lion also causing many fatalities. [1] Footage released in early 2014 of a bull elephant attacking a tourist's car in Kruger National Park, South Africa demonstrated the continued threat these animals cause. [2] Tougher protection would result in higher numbers of these animals which increases the risk to human lives. [1] Animal Danger 'Most Dangerous Animals' [2] Withnall, A. 'Rampaging bull elephant flips over British tourist car in Kruger Park' Instruction: Given the two documents, generate a question where both documents could be valid evidence. The question should be concise, and should not copy verbatim from the documents. It should be a Yes/No question where one document would support the "Yes" answer and the other would support "No". Question: Should we encourage more intense protection of endangered species? Document 1: Allowing the sale of generic drugs will not help the plight of the developing world. Many drug companies invest substantial amounts of money, gleaned from the sale of profitable dugs in the developed world, into researching treatments for the developing world. Without the revenues available from patent-protected drug sales, companies' profits will fall, precipitating a reduction in pro bono giving and research. Allowing the production of generic drugs will thus in the long run hurt the developing world. Document 2: Allowing production of generic drugs saves lives, particularly in the developing world Many developing countries are fraught with terrible disease. Much of Africa and Asia are devastated by malaria, and in many parts of Africa AIDS is a horrendous scourge, infecting large percentages of many countries populations. For example, in Swaziland, 26% of the adult population is infected with the virus1. In light of these obscenely high infection rates, African governments have sought to find means of acquiring enough drugs to treat their ailing populations. The producers of the major AIDS medications do donate substantial amounts of drugs to stricken countries, yet at the same time they charge ruinously high prices for that which they do sell, leading to serious shortages in countries that cannot afford them. The denial of the right to produce or acquire generic drugs is effectively a death sentence to people in these countries. With generic drugs freely available on the market, the access to such drugs would be facilitated far more readily and cheaply; prices would be pushed down to market levels and African governments would be able to stand a chance of providing the requisite care to their people2. Under the current system attempts by governments to access generic drugs can be met by denials of free treatments, leading to even further suffering. There is no ethical justification to allow pharmaceutical companies to charge artificially high prices for drugs that save lives. Furthermore, many firms that develop and patent drugs do not share them, nor do they act upon them themselves due to their unprofitability. This has been the case with various treatments for malaria, which affects the developing world almost exclusively, thus limiting the market to customers with little money to pay for the drugs3. The result is patents and viable treatments sitting on shelves, effectively gathering dust within company records, when they could be used to save lives. But when there is no profit there is no production. Allowing the production of generic drugs is to allow justice to be done in the developing world, saving lives and ending human suffering. 1 United Nations. 2006. "Country Program Outline for Swaziland, 2006-2010". United Nations Development Program. Available: 2 Mercer, Illana. 2001. "Patent Wrongs". Mises Daily. Available: 3 Boseley, Sarah. 2006. "Rich Countries 'Blocking Cheap Drugs for Developing World'". The Guardian. Available: Instruction: Given the two documents, generate a question where both documents could be valid evidence. The question should be concise, and should not copy verbatim from the documents. It should be a Yes/No question where one document would support the "Yes" answer and the other would support "No". Question: Should we allow the production of generic drugs? Document 1: [Doc 1] Document 2: [Doc 2] Instruction: Given the two documents, generate a question where both documents could be valid evidence. The question should be concise, and should not copy verbatim from the documents. It should be a Yes/No question where one document would support the "Yes" answer and the other would support "No". Question:
Perspective Generation	Convert the question into a statement without adding any extra information. Then also generate the negation of this statement. Question: [Question]

Table 8: The prompt we use for generating data for Arguana. [Doc 1], [Doc 2], and [Question] are substituted with the actual documents and questions during generation. The question is first generated with the prompt in the first row, and then perspectives are generated with the prompt in the second row based on the question.

Type	Prompt
Yes/No	Given the question and a positive statement, generate a statement of opposite stance with respect to the question. Positive Statement: [Statement] Question: [Question] Negative Statement:
Multi	Given the answers and the question, generate statements that are followed by a newline character. The statement should closely follow the format of the question, and the statement should only differ by the segment where the answer is. Question: [Question] Answers: [Answers] Statements:

Table 9: The prompts we use for generating data for Kialo. Different prompts are used according to the data type (whether the question allows more than two perspectives). [Statement],[Question], and [Answers] are substituted with the statement, the question, and the answers provided by Kialo.com respectively during generation. Note that a list of answers is only provided for the questions with more than two perspectives.

Dataset	Original Data (Provided by the Source)	Generated Data
Arguana	**Document 1**: People are given too much choice, which makes them less happy. Advertising leads to many people being overwhelmed by the endless need to decide between competing demands on their attention - this is known as the tyranny of choice or choice overload. ... **Document 2**: People are unhappy because they can't have everything, not because they are given too much choice and find it stressful. In fact, advertisements play a crucial role in ensuring ...	**Question**: Does an increase in choice lead to unhappiness and stress? **Supporting Perspective**: An increase in choice leads to unhappiness and stress. **Opposing Perspective**: An increase in choice does not lead to unhappiness and stress.
Kialo (Yes/No Question)	**Question**: Will ChatGPT do more harm than good? **Supporting Perspective**: ChatGPT will do more harm than good.	**Opposing Perspective**: ChatGPT will do more good than harm.
Kialo (More than two perspectives)	**Question**: What's the best Twitter/X alternative? **Answer1**: Facebook **Answer2**: Kialo **Answer3**: Instagram **Answer4**: Reddit	**Perspective 1**: The best Twitter/X alternative is Facebook. **Perspective 2**: The best Twitter/X alternative is Kialo. **Perspective 3**: The best Twitter/X alternative is Instagram. **Perspective 4**: The best Twitter/X alternative is Reddit. ...
OpinionQA	**Original Question**: Would having a gun in your household make you feel ['Safer than you feel without a gun in your household', 'Less safe than you feel without a gun in your household', 'No more or less safe', 'Refused']	**Generated Question**: Do you feel safer with a gun in your household? **Perspective 1**: Having a gun in the household increases my sense of safety. **Perspective 2**: Having a gun in the household does not increase my sense of safety.

Table 10: An example of data provided by the original source vs. generated. Data in the "Original Data" column is either provided by the source dataset (Arguana and OpinionQA) or Kialo.com. Kialo.com provide different information for different types of questions (whether the question is a yes/no question or not). Questions in **blue** and perspectives in **red** are the ones we include in BERDS.

Setting	Prompt
Question Filtering	Determine if the question is asking about personal experience or information. Reply with only "Yes" or "No". Only answer "Yes" if (1) The question is asking about if you have ever done something or if something has happened to you. (2) The question is asking about facts about you, for example, if you own something, or how you describe yourself. (3) The question is asking about something that has happened to your local community, or the current status of your local community. You should answer "No" if the question is asking about your opinion, such as if you are worried about something or what you think about a certain subject. Question: [Question] Answer:
Data Generation	Convert the following survey question into a Yes/No question, with the first and the last choice in the brackets as the two answers. Generate the question first, and then convert the question into a positive and a negative statement that supports and oppose the question respectively. The statements should be concise and not contain any additional information that is not in the question. The statements should also be fluent and grammatically correct. Question: [Question] Converted question:

Table 11: The prompt we use for generating data for OpinionQA. [Question] is substituted with the original question, which comes with a list of options, in the OpinionQA dataset during generation. We first filter out data that are not related to personal experience or facts about the survey taker. Then we generate questions and perspectives based on the original question provided by OpinionQA.

Generate diverse perspectives that answer the given question. Please answer with a JSON object, where each field is a perspective. Do not generate similar perspectives; they should be distinct. The keys in the JSON object should be meaningful, and each word in the keys should be separated by whitespace. Question: [Question]

Table 12: The input prompt for generating diverse perspectives for query expansion. [Question] is substituted with the question of each example. The output would be a JSON object containing the perspectives.

Dataset	Question	GPT-4-Generated Perpsectives
Arguana	**Question**: Does an increase in choice lead to unhappiness and stress?	**Increased_Choice_Leads_to_Paralysis**: An increase in choice may lead to feelings of unhappiness and stress, due to the decision paralysis phenomena. With too many options, individuals may struggle to make a choice and fear making the wrong decision. **Choice_Overload_Stress**: When faced with numerous choices, individuals may feel overwhelmed, raising their stress levels. Too many options may make decision-making processes lengthy and taxing, escalating stress and leading to unhappiness. **Regret_Fear_and_Happiness_Decrease**: Increased choice can potentially lead to decreased happiness and increased stress, as individuals may worry about their decisions, regret not choosing different options, or fear missing out on potentially better choices.
Kialo	**Question**: Will ChatGPT do more harm than good?	**Artificial_Intelligence_Opportunity**: ChatGPT, like other AI models, can have immense potential and bring about substantial benefits. By enabling natural language interaction, it can be used in areas such as education, healthcare, customer service, and more, greatly enhancing efficiency and user experience. The potential good that ChatGPT can contribute far outweighs the potential harm, assuming that the proper controls and ethical guidelines are put in place. **Miscommunication_Risk**: While ChatGPT is designed to understand and generate human-like text, there's a risk of miscommunication. The AI model may not fully grasp subtle nuances, cultural contexts, or emotions which could lead to misunderstanding or even harm, especially in sensitive conversation scenarios. **Data_Privacy_Concern**: There's a potential for misuse pertaining to data privacy with ChatGPT. It's trained on incredibly large datasets, and there might be instances where personal or sensitive data might be leaked or misused, posing a serious threat to privacy.
OpinionQA	**Question**: Do you feel safer with a gun in your household?	**Second_amendment_advocate_perspective**: Yes, I do feel safer with a gun in the household. It provides a means of defending my family, property and myself from potential offenders or burglars. **Child_safety_concerned_perspective**: No, I don't feel safer with a gun in the home. Especially when there are kids around, the risk of accidents due to negligence is significant and unnerving. **Non_violence_promoter_perspective**: No, owning a gun insinuates the readiness for violence, which creates an atmosphere of constant threat and insecurity. I believe in resolving conflicts peacefully rather than aggressively.

Table 13: Examples of diverse perspectives generated by GPT-4, later used for query expansion in Section 5. We only present the first three generated perspectives due to the page limit.

Prompt	<lim_start|>system You are a helpful assistant that can only answer with "Yes" or "No". Your goal is to determine if the provided document support the provided statement. Do NOT provide any explanation for your choice. <lim_end|> [Optional 1-shot example] <lim_start|>user Document: [Document] Statement: [Statement] Instruction: Does the document support the statement? Answer with only "Yes" or "No". Only consider the information provided in the document, and do not infer any additional information. The document is a snippet taken from the Internet, and might exhibit support for or opposition against the statement. If the document opposes the statement, the answer is "No". If the document does not contain relevant information about the statement, the answer is also "No". The answer is "Yes" only if the document supports the statement, either explicitly or implicitly. Think carefully if the document actually supports the statement, or if there is just some superficial textual overlap between the two. Answer: <lim_end|>
One-Shot Example	<lim_start|>user Document for Example 1: Why Prop. 19 is so important | rescue truth. The United States took a stab at alcohol prohibition from 1920-1933 via the United States Constitution. Almost immediately following the ratification of the Eighteenth Amendment, which outlawed alcohol, speakeasy clubs sprung up all over, effectively creating a black market and handing alcohol sales over to organized crime. During this time, our government poisoned industrial alcohol in an attempt to curb usage by scaring the American public, causing an estimated 10,000 deaths.[17] Because of the illicit drug market created by prohibition, the government could not regulate the production of alcohol, and sometimes people became ill after drinking bootlegged whiskies. The current Statement for Example 1: Prohibition is an effective method to curb drug usage. Instruction for Example 1: Does the document support the statement? Answer with only "Yes" or "No". Only consider the information provided in the document, and do not infer any additional information. The document is a snippet taken from the Internet, and might exhibit support for or opposition against the statement. If the document opposes the statement, the answer is "No". If the document does not contain relevant information about the statement, the answer is also "No". The answer is "Yes" only if the document supports the statement, either explicitly or implicitly. Think carefully if the document actually supports the statement, or if there is just some superficial textual overlap between the two. Answer for Example 1: <lim_end|> <lim_start|>assistant No <lim_end|>

Table 14: The input prompt for perspective detection. We provide an optional one-shot example for some models that do not do well zero-shot. [Document] and [Statement] are substituted with the document and perspective in each example respectively.

Question: [Question] Perspective 1: [P1] Perspective 2: [P2] Instruction: Given the question, one supporting and one opposing perspectives, which perspective is supporting the question? Answer with only “Perspective 1” or “Perspective 2”. Answer:

Table 15: The input prompt for deciding which of the perspectives supports or opposes the question. We only consider examples with two perspectives. We replace [Question] with the actual question, and [P1], [P2] with the two perspectives when prompting.

Corpus	Model	Datasets							
		Arguana		Kialo		OpinionQA		Average	
		Prec.	MRec.	Prec.	MRec.	Prec.	MRec.	Prec.	MRec.
Wiki	BM25	19.81	19.33	20.58	16.67	9.13	5.67	16.51	13.89
	DPR	17.35	14.67	18.17	13.57	12.03	6.35	15.85	11.53
	CONTRIEVER	32.37	31.07	30.71	26.10	21.50	13.83	28.19	23.67
	TART	34.53	35.47	31.53	26.53	23.40	15.07	29.82	25.69
	NV-Embed-v2	36.87	36.67	35.89	32.13	26.52	14.53	33.09	27.78
Sphere	BM25	42.44	43.20	41.54	37.34	44.09	33.67	42.69	38.07
	DPR	44.09	33.67	13.51	12.40	8.16	6.12	21.92	17.40
	CONTRIEVER	55.89	55.07	53.89	43.28	48.61	38.66	52.80	45.67
	TART	**65.44**	**59.87**	56.45	44.80	**63.08**	**41.33**	**61.66**	48.67
Google Search Output	BM25	33.0	36.40	34.92	38.50	33.81	29.25	33.91	34.72
	CONTRIEVER	48.65	44.53	47.95	44.32	42.73	35.49	46.44	41.45
	TART	51.43	45.33	49.99	43.20	45.69	35.07	49.04	41.20
	NV-Embed-v2	59.17	55.87	**61.61**	**51.60**	50.32	40.13	57.03	**49.20**

Table 16: Performance on the test split of BERDS, considering top 10 retrieved documents.

Corpus	Model	Avg. size (↓)	% within top 100 (↑)
Wiki	BM25	29.8	41.5
	DPR	30.9	36.7
	CONT.	24.7	55.6
Sphere	BM25	20.8	77.3
	DPR	37.4	48.9
	CONT.	17.8	83.1
Google	BM25	21.0	72.1
	CONT.	18.2	76.2

Table 17: The average size of the document set required to cover all perspectives with base retrievers. The average size is only computed over the examples where the sets of the top 100 documents contain all perspectives (MRECALL @ 100 = 1), and we report % of data that satisfies this (MRECALL @ 100) in the last column.

C Additional Results

C.1 Results for Top 10 Documents

We report results for retrieving $k = 10$ documents in Table 16. Comparing the results with those in Table 3 shows that including more documents leads to higher diversity (MRECALL) at the cost of PRECISION due to the reduced relevance of lower-ranked documents. Similar trends hold for $k = 10$, as CONTRIEVER is still the best retriever in terms of diversity, and retrievers obtain more diverse documents on Sphere, followed by Google Search API outputs and Wikipedia.

C.2 How many documents should one retrieve to cover diverse perspectives?

In Table 3, we have reported the performances of the retrieval outputs consisting of five documents and found that the model outputs do not cover diverse perspectives. If we consider up to the top 100 documents, can we cover diverse perspectives? What is the average number of documents needed to be retrieved to cover all perspectives? Table 17 reports the size of the retrieved document set needed to cover all the perspectives with base retrievers. We find that the best setting (CONTRIVER with Sphere) can find all the perspectives within the top 100 document set 83.1% of the time, and for those examples, the average number of documents required to cover all the perspectives is 17.8. Other settings show worse performance, mostly reflecting performance reported in Table 3.