PerSphere: A Comprehensive Framework for Multi-Faceted Perspective Retrieval and Summarization

Yun Luo[1], **Yingjie Li**[1], **Xiangkun Hu**[2]*; **Qinglin Qi**[3],
Fang Guo[1], **Qipeng Guo**[4], **Zheng Zhang**[5], **Yue Zhang**[1,6] ✉†
[1]School of Engineering, Westlake University [2]Amazon AWS AI.
[3]Sichuan University. [4]Shanghai AI Lab. [5]New York University, Shanghai.
[6] Institute of Advanced Technology, Westlake Institute for Advanced Study
{luoyun, zhangyue}@westlake.edu.cn

Abstract

As online platforms and recommendation algorithms evolve, people are increasingly trapped in echo chambers, leading to biased understandings of various issues. To combat this issue, we have introduced PerSphere, a benchmark designed to facilitate multi-faceted perspective retrieval and summarization, thus breaking free from these information silos. For each query within PerSphere, there are two opposing claims, each supported by distinct, non-overlapping perspectives drawn from one or more documents. Our goal is to accurately summarize these documents, aligning the summaries with the respective claims and their underlying perspectives. This task is structured as a two-step end-to-end pipeline that includes comprehensive document retrieval and multi-faceted summarization. Furthermore, we propose a set of metrics to evaluate the comprehensiveness of the retrieval and summarization content. Experimental results on various counterparts for the pipeline show that recent models struggle with such a complex task. Analysis shows that the main challenge lies in long context and perspective extraction, and we propose a simple but effective multi-agent summarization system, offering a promising solution to enhance performance on PerSphere.

1 Introduction

The rapid expansion of social platforms and personalized recommendation systems on the web has significantly increased the risk of confining individuals within echo chambers (Nguyen, 2020), limiting their exposure to diverse perspectives and encouraging the formation of homogeneous user groups that perpetuate similar narratives (Cinelli et al., 2021; Nyhan et al., 2023; Lo et al., 2021). It can lead to biased or incomplete viewpoints, opinion polarization, or the spread of misinformation (Del Vicario et al., 2016; Falkenberg et al., 2022; Van Der Linden, 2022). In real-world scenarios, users may not seek a universally accepted "ground truth". Instead, they look for a comprehensive understanding and evidence on controversial issues (Wang and Ling, 2016). For example, compared to a simple answer, a detailed and balanced argumentative summary of the topic can be more useful. To this end, multi-faceted perspective retrieval and summarization (MURS) could help individuals access comprehensive and critical information (Turan et al., 2019).

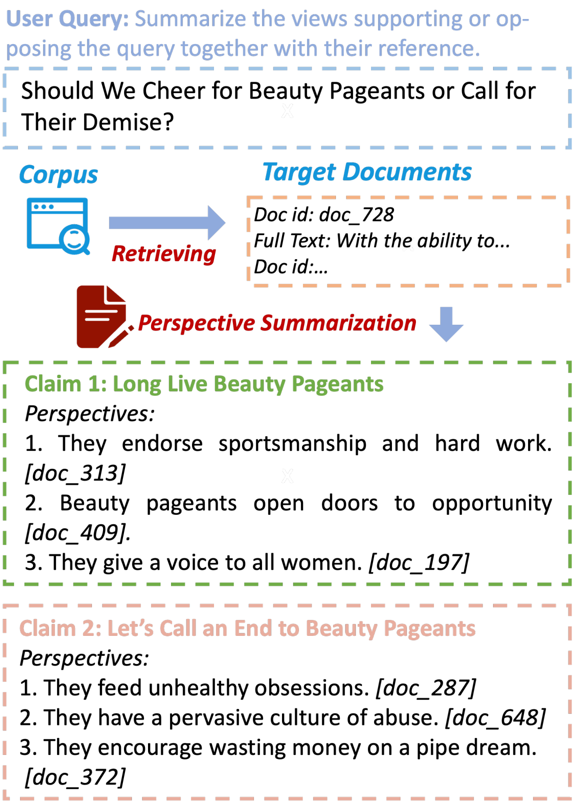

User Query: Summarize the views supporting or opposing the query together with their reference.
Should We Cheer for Beauty Pageants or Call for Their Demise?
Corpus
Target Documents
Retrieving
Doc id: doc_728
Full Text: With the ability to...
Doc id:...
Perspective Summarization
Claim 1: Long Live Beauty Pageants
Perspectives:
1. They endorse sportsmanship and hard work. [doc_313]
2. Beauty pageants open doors to opportunity [doc_409].
3. They give a voice to all women. [doc_197]
Claim 2: Let's Call an End to Beauty Pageants
Perspectives:
1. They feed unhealthy obsessions. [doc_287]
2. They have a pervasive culture of abuse. [doc_648]
3. They encourage wasting money on a pipe dream. [doc_372]

Figure 1: Multi-faceted perspective retrieval and summarization, where the system outputs a summary of two controversial claims and several non-overlapping perspectives together with corresponding references.

There has been existing work on argumentative

[*] The work does not relate to the author's position at Amazon.

[†] Corresponding author.

summarization (Ajjour et al., 2019; Syed et al., 2023; Li et al., 2024). However, such work does not consider the comprehensiveness of the generated content, in which significant challenges exist. First, retrieval-augmented generation (RAG) methods focus on retrieving documents relevant to the query, but do not explicitly consider the comprehensiveness of perspectives across the retrieved data. As shown in the Figure 1, the user may expect to know about whether we should cheer for beauty pageants, and the retriever might successively search for documents on the impact of beauty pageants on individual unhealthy obsessions, but omit other documents due to a lower match score. Second, summarizing the retrieved documents into concise, one-sentence perspectives that capture key points without overlapping information presents a considerable challenge (Friedman et al., 2021). Third, generating perspectives or answers with corresponding references is also difficult, as demonstrated by previous studies (Gao et al., 2023a; Huang and Chang, 2024).

We benchmark the proposed task MURS by presenting PerSphere, a dataset composed of 1,064 instances, consisting of a specific query with two controversial claims. Each claim is supported by various perspectives drawn from a document set. The objective is to provide a comprehensive summary of perspectives by mining the set of documents in response to a given query, in the form of a structured summarization framework that includes claims, perspectives, and document references. To assess task performance, we propose a specific set of evaluation metrics such as Recall@k, Cover@k for retrieval, and GPT-4 score for summarization using a specifically designed prompt. Meta Evaluations confirm the effectiveness of the GPT-4 score for summarization.

We use PerSphere to evaluate both open-source and closed-source large language models (LLMs), including LLaMA-3.1-8B-Inst, LLaMA-3.1-70B-Inst, Claude-3-Sonnet, and GPT-4-Turbo. Results using the standard RAG pipeline reveal that the models encounter challenges in retrieval and summarization within PerSphere. Analysis shows that the salient challenge lies in long context and extracting one-sentence perspectives. Consequently, we propose HierSphere, a multi-agent summarization system to enhance the multi-faceted summarization. This approach initially employs agents to generate local summaries from different sets of the retrieved documents. Subsequently, an editorial agent merges these summaries further and refines the perspectives focusing on the central themes. Results show that HierSphere achieves stronger performance because of the reduction of the long context and refinement of the perspectives. We will release our dataset and codes on the link[1].

2 Related Work

2.1 Argument Mining

Extensive research has been dedicated to formalizing argumentative structures from free text, encompassing various sub-tasks such as claim identification (Rinott et al., 2015; Levy et al., 2018), argument structure extraction (Luo et al., 2023b), argument classification (Stab et al., 2018b; Li and Caragea, 2023), evidence identification (Shnarch et al., 2018; Ein-Dor et al., 2020), argument summarization and clustering (Ajjour et al., 2019; Syed et al., 2023), and key point analysis (Friedman et al., 2021). Given the widespread application of retrieval from websites and document corpora, studies have also attempted to construct perspective retrieval or extraction systems. Stab et al. (2018a) presented an argument retrieval system capable of retrieving sequential arguments for any given controversial topic. Reimer et al. (2023) proposed a reranking approach to improve retrieval effectiveness for non-factual comparative queries by considering the stance of the object relative to the comparison object. These systems primarily focus on the relevance of retrieved documents to a specific claim.

Similarly, Li et al. (2024) proposed an end-to-end summarization system (ArgSum) to prepare argumentative essays for debates. Unlike our study, which emphasizes the comprehensiveness of perspectives in a specific document set, their work focuses on the convincingness of the summarization of a specific claim. Another similar work is Perspectrum (Chen et al., 2019), which aims to retrieve and cluster relevant perspectives to a claim from a given perspective pool and retrieve documents to support each perspective. The salient difference is that Perspectrum assumes that perspectives can be extracted before the query or claim is given, which is not realistic in applications.

2.2 Retrieving-augmented Generation

Large language models (LLMs) demonstrate impressive capabilities, yet face significant challenges

[1] https://github.com/LuoXiaoHeics/PerSphere

in practical applications (Gao et al., 2023b) such as hallucinations, stubbornness to internal knowledge (Yan et al., 2024, 2025), and catastrophic forgetting for knowledge injection (Luo et al., 2023a). Retrieval-augmented generation (RAG) addresses these issues by dynamically retrieving information from external knowledge sources and using it as references to formulate answers (Gao et al., 2023b). For instance, in healthcare, RAG systems aid evidence-based decision-making and potentially improve patient care by retrieving and summarizing relevant studies, clinical trials, and treatment guidelines (Wang et al., 2024). Some RAG systems, such as Raptor (Sarthi et al., 2024) or GraphRAG (Edge et al., 2024), were also proposed to enhance holistic understanding of the overall document context, but only focused on the QA tasks. Our task naturally fits an RAG pipeline, where the retriever aims to retrieve relevant documents and the generator responds with the retrieved content. However, different from a general RAG system, we focus on the comprehensiveness of the perspectives in a document set and multi-faceted summarization.

3 PerSphere

Before delving into the specifics of data construction, we first outline the task formulation of our multi-faceted perspective retrieval and summarization task (Section 3.1). Our dataset PerSphere is composed of two subset, each covering a different level of difficulty: (1) Theperspective, sourced from the editorial website THEPERSPECTIVE[2] (Section 3.2); (2) Perspectrumx, which is developed based on the dataset Perspectrum (Section 3.3). The data statistics are summarized in Section 3.4. Finally, a set of evaluation metrics is proposed for our task (Section 3.5).

3.1 Task Formulation

Given a document set D, and a query q^i, our pipeline contains two steps: 1) comprehensive document retrieval, where we aim to retrieve k comprehensive documents from D; 2) multi-faceted summarization, where we summarize two controversial claims c_0^i and c_1^i, and their perspectives p^i to the query q^i. Denote related documents in D to the query q^i as D^i, and the corresponding golden perspectives are denoted as p^i. Each perspective p_j^i can be extracted from several documents in D^i, denoted as D_j^i ($|D_j^i| >= 1$). Any perspective pair p_j^i and p_k^i in p^i are not overlapped with each other, and each perspective conveys a viewpoint to the claim c^i. The task can be formulated as (we omit the instance number i for simplicity)

$$q \rightarrow c_0 : \{p_{0,j}, [D_{0,j}]\};\ c_1 : \{p_{1,j}, [D_{1,j}]\}. \quad (1)$$

where $0, 1$ refers to two controversial claims, and [] refers to corresponding references. Data instances are shown in Appendix G.

Table 1: The statistics of our datasets. '#' refers to the data number, and 'docs' refers to the abbreviation of documents.

Dataset	Statistics	Value
Theperspective	# of the queries	185
	# of the perspectives	1,103
	supporting perspectives	552
	undermining perspectives	551
	average perspectives	5.96
	# of docs	4,107
	average length of docs	131.2
Perspectrumx	# of the queries	878
	# of the perspectives	4,493
	supporting perspectives	2,488
	undermining perspectives	2,005
	average perspectives	5.11
	# of docs	8,092
	average length of docs	168.5

3.2 Theperspective

Firstly, we collect data from the THEPERSPECTIVE website, which hosts numerous editorial articles including the topics of Lives, Sports, Politics, Entertainment and Technology. Each article contains a query title with two controversial claims, supported by several perspectives. For each perspective, an evidence paragraph is provided to discuss and support the corresponding perspective. We collect 221 articles from the website up to July 15th, 2024, and use the BeautifulSoup package[3] to extract topics, claims, perspectives, and evidence documents. After filtering out unstructured data, we obtain a final dataset of 185 effective entries. Each entry contains two claims, with each claim supported by 2 to 4 perspectives. It's important to note that in the THEPERSPECTIVE dataset, each perspective corresponds to only one evidence document, which simplifies the task of retrieving relevant documents.

Given that the perspectives in the documents may be incomplete sentences or phrases lacking

[2] https://www.theperspective.com/

[3] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

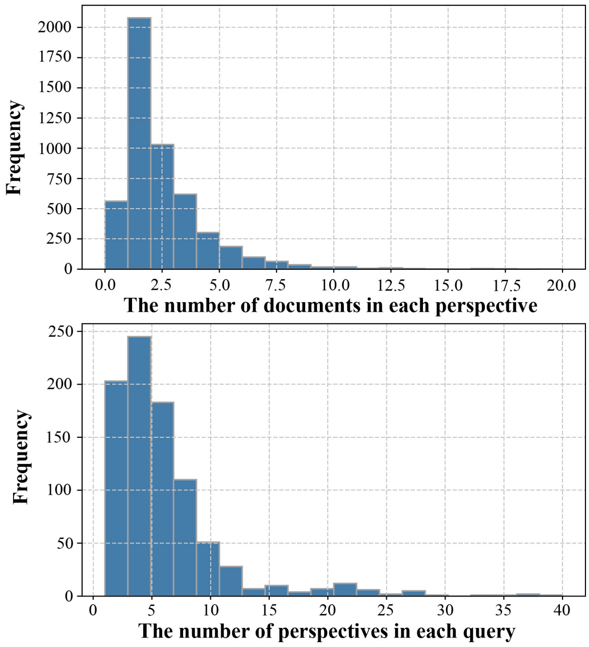


Figure 2: The frequency of the number of documents and perspectives in Perspectrumx.

completed semantic meaning, we employ GPT-4-Turbo to review and complete these perspectives. The detailed prompt used for this process is provided in the Appendix A. To further enhance document diversity, we then add documents from Perspectrum to the Theperspective corpus that are unrelated to the queries in Theperspective (the details are also shown in Appendix A). This process expands the document set from 1,103 to 4,107 documents, significantly enhancing the diversity of the document collection.

3.3 Perspectrumx

Different from Theperspective where only one evidence document supports each perspective, we construct a more challenging set with multiple documents. Specifically, we construct our *Perspectrumx* based on Perspectrum (Chen et al., 2019), which includes claims, perspectives, and evidence derived from online debate data. Since there are no explicit queries in Perspectrum, we adopt the 'claims' in Perspectrum as our queries and randomly select one 'perspective' from each cluster (with semantically equivalent meaning) in Perspectrum as our perspectives. We exclude queries that lack perspectives and perspectives that do not have associated evidence documents. The evidence documents are used directly. Our claims are directly generated following a specific template: "We support/undermine the argument that {query}". We show the details in Appendix H for further explanation.

Prompt to evaluate summarization

Please serve as an impartial judge and assess the quality of a summarization performed by an AI assistant. You should strictly evaluate it with the golden summarization but not your own knowledge or assumption. We will provide both the ideal summarization ("golden summarization") and the AI's response.
Your evaluation should adhere to the following requirements:
1. The perspectives generated should accurately reflect the claims.
2. The references provided should support the corresponding perspectives.
3. Each perspective should be distinct and free from irrelevant information or overlap with others.
4. The perspectives should comprehensively cover the content of the golden summarization.
5. The content of the response should not be fabricated or expanded; it should not include anything that does not appear in the golden summarization.
6. The sequence of perspectives in the response does not need to match that of the golden summarization.
7. The response is provided in an XML file, and the format difference is not a critical criterion.
Considering these factors, please begin your evaluation with a brief explanation, aiming for maximum objectivity. After providing your explanation, rate the response on a scale from 1 to 10, using this format: "Rating: [[number]]" (for example, "Rating: [[5]]").
Golden Summarization: {golden}
Response: {response}

Figure 3: Evaluation prompt for summarization.

3.4 Statistics

Data statistics are shown in Table 1. In summary, our dataset comprises a total of 1,063 instances, with 185 instances in Theperspective and 878 instances in Perspectrumx. The key distinction between the two subsets is that Theperspective includes just one evidence document per perspective, whereas Perspectrumx includes multiple evidence documents per perspective. The data distribution of Perspectrumx is illustrated in Figure 2, which exhibits a Poisson-like distribution. This distribution indicates significant variability in the number of evidence documents associated with each perspective, as well as in the number of perspectives available for each query. Such variability poses a considerable challenge in our task. Merely extracting relevant documents without ensuring comprehensiveness would be insufficient in addressing this scenario.

3.5 Evaluation Metric

Retrieval. For the retrieval results, we propose to use the metric Recall@k to evaluate the relevance of the retrieved documents, which is calculated as

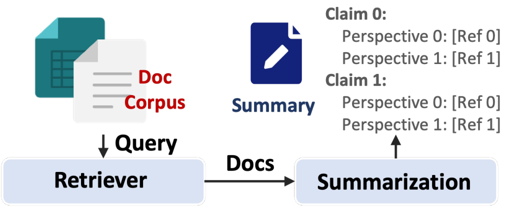


Figure 4: The RAG pipeline of our multi-faceted perspective retrieval and summarization.

follows:

$$\text{Recall@k} = \frac{1}{N}\sum_{i}^{N}\frac{|\{d_m \in D^i\} \cap \{d_m \in T_k\}|}{|D^i|} \quad (2)$$

where D^i is the set of relevant documents to query q_i, T_k is the set of top k retrieved documents, and d_m represents an individual document. We also propose the coverage of the perspectives in the retrieved documents, denoted as Cover@k:

$$\frac{1}{N}\sum_{i}^{N}\frac{|\{p_j^i \leftarrow d_m\} \cap \{d_m \in D_j^i\} \cap \{d_m \in T_k\}|}{|p^i|}. \quad (3)$$

where $p_j^i \leftarrow d_m$ refers to the perspectives can be covered in documents d_m, and p^i are the perspectives for the query q^i.

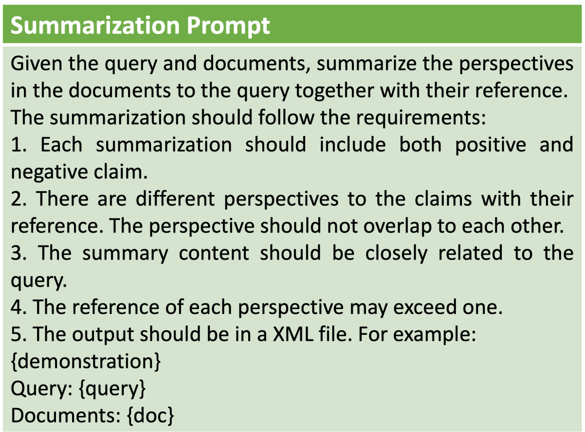
Summarization. Inspired by MTBench (Zheng et al., 2023; Hu et al., 2024), which uses LLMs as evaluators, we propose a GPT-4 evaluation score for the multi-faceted summarization. In particular, we establish specific criteria to assess the performance of the generated content. For example, each perspective should be distinct and free from irrelevant information or overlap with others. The details are outlined in the Figure 3.

4 Standard RAG Pipeline

Our task can fit in the RAG pipeline as illustrated in Figure 4. To evaluate the performance of current retrieval models and LLMs in the multi-faceted summarization task, we carry out experiments in various retrieval baselines (Section 4.1) and LLMs for multi-faceted summarization (Section 4.2).

4.1 Retriever

Following Ajith et al. (2024), we evaluate the baselines such as sparse retrieval BM25 (Robertson and Zaragoza, 2009) as well as several state-of-the-art dense retrieval models including E5 (Wang et al., 2022), GTR-T5 (Ni et al., 2022), text-embedding-ada-002[4], and GritLM (Muennighoff et al.). Details are shown in Appendix E. Specifically, we use the open-source models from Huggingface, including E5-large[5], GTR-T5-large[6], and GritLM-7B[7] for implementation. The dense retrievals are based on the Faiss package[8].



Summarization Prompt

Given the query and documents, summarize the perspectives in the documents to the query together with their reference. The summarization should follow the requirements:
1. Each summarization should include both positive and negative claim.
2. There are different perspectives to the claims with their reference. The perspective should not overlap to each other.
3. The summary content should be closely related to the query.
4. The reference of each perspective may exceed one.
5. The output should be in a XML file. For example:
{demonstration}
Query: {query}
Documents: {doc}

Figure 5: Summarization prompt, where $\{\cdot\}$ refers to the input content, and "demonstration" refers to an XML sample (abbreviated here to save space).

4.2 Generator

To evaluate the performance of current LLMs in perspective summarization, we design a specific prompt to generate summaries using retrieved documents within an XML file structure. To ensure consistent output formatting, we also provide a demonstration to guide the response. The prompt is shown in Figure 5. We evaluate the instruction-tuned models such as LLaMA-3.1-8B-Instruct, LLaMA-3.1-70B-Instruct (for simplicity we omit Instruct in the following description), Claude-3-Sonnet[9], and GPT-4-Turbo[10] for summarization.

4.3 Retrieval Results

We first present the retrieval results in Table 2. The recall performance in Theperspective is significantly higher than in Perspectrumx, indicating a relatively lower difficulty of Theperspective. For example, GritLM achieves a recall@20 of 96.77% in Theperspective, which is 26.19% higher than in Perspectrumx (70.58%). Additionally, GritLM consistently outperforms other retrieval methods in

[4] https://platform.openai.com/docs/guides /embeddings/embedding-models

[5] https://huggingface.co/intfloat/e5-large-v2

[6] https://huggingface.co/sentence-transformers/gtr-t5-large

[7] https://huggingface.co/GritLM/GritLM-7B

[8] https://github.com/facebookresearch/faiss

[9] We adopt the Claude-3 version *claude-3-sonnet-20240229*

[10] We adopt the GPT-4-Turbo version *gpt-4-0125-preview*

Table 2: The performance of the retrieval based on different retrievers. Recall@k refers to the ratio of relevant documents in k retrieved documents to the golden documents. Cover@k refers to the coverage values of the retrieved documents to the perspectives. In Theperspective, Recall@k equals Cover@k.

	Theperspective		**Perspectrumx**			
Metrics	Recall@10	Recall@20	Recall@10	Cover@10	Recall@20	Cover@20
BM25	72.37	82.35	44.71	56.20	56.27	64.43
E5-large	78.58	88.89	49.45	60.88	61.18	70.17
GTR-large	85.66	94.80	53.83	64.52	65.68	72.98
Ada-002	86.47	95.34	53.43	64.43	68.80	74.16
GritLM	**90.50**	**96.77**	**58.21**	**68.71**	**70.58**	**77.01**

Table 3: The results of the summarization based on the evaluation of GPT-4-Turbo, where the performance Overall@k is score between 1-10, and a large score refers to more consistent performance to the golden summarization. 'Golden' refers to the summarization performance given the golden relevant documents.

		Thepersective		**Perspectrumx**	
Retriever	Sumarization	Overall@10	Overall@20	Overall@10	Overall@20
BM25	LLaMA-3.1-8B	6.01	6.22	4.37	4.40
	LLaMA-3.1-70B	7.16	7.35	4.95	4.97
	GPT-4-Turbo	7.52	7.74	5.66	5.49
	Claude-3-Sonnet	7.34	7.77	5.64	5.43
GritLM	LLaMA-3.1-8B	6.81	6.41	4.95	4.90
	LLaMA-3.1-70B	7.68	7.54	5.23	5.11
	GPT-4-Turbo	**7.99**	**8.16**	**6.20**	6.04
	Claude-3-Sonnet	7.91	8.05	6.16	**6.25**
Golden	LLaMA-3.1-8B	8.08		6.10	
	LLaMA-3.1-70B	8.58		6.63	
	GPT-4-Turbo	8.75		7.33	
	Claude-3-Sonnet	8.80		7.28	

Theperspective and Perspectrumx, whereas BM25 exhibits the lowest performance. We also observe that as k increases, both Recall and Cover improve, such as an increase from 90.50% to 96.77% when k rises from 10 to 20. Finally, the performance of Recall@k positively correlates with Cover@k, which indicates a stronger retrieval method can generally achieve more comprehensive results. We further report the Recall@100 and Cover@100 in Appendix F, which can be applied to the method of re-ranking.

4.4 Summarization Results

The summarization results are shown in Table 3. Several observations are summarized as follows:

Multiple evidence documents increase the difficulty in multi-faceted retrieval and summarization. The summarization performance in Theperspective is significantly higher than those in the Perspectrumx, indicating that with only one reference document for each perspective, the task would be relatively simple. For example, with the golden documents, Claude-3-Sonnet could achieve an 8.8 score in Theperspective, but only 7.28 in Perspectrumx. The performance using retrieval methods and other LLMs is similar.

Better retriever and LLMs lead to enhanced summarization performance. For example, by using GritLM, LLaMA-3.1-8B achieves a score of 6.81 in Theperspective, which is 0.8 points higher than the score obtained using BM25. However, there are still large margins to the summarization performance in using the golden evidence documents. The score in Claude-3-Sonnet is 7.28 using golden documents, which is 1.63 higher than that using GritLM, and 1.85 higher than BM25. As for the LLMs, Claude-3-Sonnet and GPT-4-Turbo achieve the most promising performance, whereas LLaMA-3.1-8B achieves the weakest performance. It demonstrates that the summarization model plays a significant role in the multi-faceted perspective retrieval and summarization task.

The retrieval of more documents does not consistently enhance summarization performance. This phenomenon is most prominently observed in Perspectrumx. Claude-3-Sonnet achieves a higher score of 5.64 when using 10 documents retrieved by BM25, compared to 5.43 with 20 documents. This discrepancy underscores the challenges of multi-faceted perspective summarization. In further experiments, we utilize Claude-3-Sonnet and LLaMA-3.1-70B to summarize collections of 100 documents retrieved by GritLM, achieving GPT-4 scores of 4.43 and 6.06, respectively. We observe that the models often incorporate irrelevant information from their internal knowledge bases and generate complex, multi-perspective content from a single perspective. These findings indicate that current models continue to struggle with processing long contexts, and that summarization based on a smaller set of retrieved documents tends to yield superior performance.

Document orders influence the performance of summarization tasks. We experiment by adjusting the order of documents in the retrieved sets to both reverse and random sequences, subsequently prompting LLMs to generate summaries. The results, as depicted in Figure 6, show that model performance generally declines when documents are presented in reverse or random order compared to the vanilla order. This suggests that current LLMs tend to lose focus midway through the documents and primarily concentrate on the information presented at the beginning. This observation aligns with findings from Liu et al. (2024), highlighting the difficulties that LLMs face when dealing with long contexts in summarization tasks.

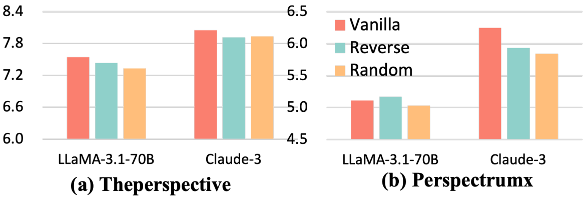
With the golden relevant documents, it is still challenging in the summarization task. The summarization performance is the most promising compared with those using the documents retrieved by the retrievers, but there is still significant space for improvement. For example, Claude-3-Sonnet can achieve a score of 8.8 in Theperspective, and GPT-4-Turbo achieves 7.33 in Perspectrumx, which represents the state-of-the-art performance in the summarization baselines.

4.5 Analysis of Perspective Extraction

Given the moderate results of Cover@k shown in Table 2, we further investigate whether perspectives can be extracted from the retrieved documents using an advanced LLM, denoted as $\mathcal{M}$. To this end, we introduce the metric Rp@k, defined by the following equation:

$$\frac{1}{N}\sum_{i}^{N}\frac{|p_j^i \Leftarrow \mathcal{M}(d_m, q_i) \cap d_m \in D_j^i \cap d_m \in T_k|}{|p^i|}, \quad (4)$$

where $\mathcal{M}(d_m, q_i)$ indicates the model's inference regarding the perspective, and '$\Leftarrow$' refers to the entailment to p_j^i. We specifically employ GPT-4-Turbo to assess the entailment of the generated perspectives against the gold standard perspectives. The prompts for perspective extraction and evaluation are detailed in Appendix B.



(a) Theperspective

(b) Perspectrumx

Figure 6: The performance with different orders of retrieved documents in GritLM@20.

Table 4: The performance of the perspective extraction based on different retrievers.

	Theperspective		Perspectrumx	
Metrics	Rp@20	Rp@10	Rp@20	Rp@10
BM25	65.36	58.05	48.71	41.76
E5-large	71.41	63.48	54.30	46.18
GTR-large	76.18	69.06	56.27	48.70
Ada-002	76.00	69.42	57.55	49.20
GritLM	**77.35**	**72.58**	**60.23**	**52.13**

Without loss of generality, we use GPT-4-Turbo as an extraction example due to its strong capabilities. The results of Rp@k are shown in Table 4. As observed, although Cover@k is notably high in retrieval methods, Rp@k lags significantly behind. For example, GritLM achieves a Cover@20 of 96.77% in Theperspective, yet its Rp@20 is only 77.35%. This discrepancy indicates that although the retrieved documents cover a wide range of perspectives, perspective extraction continues to pose a considerable challenge. Besides, we evaluate the model performance using the golden documents in Appendix D, which indicates that the models struggle to extract one-sentence perspectives, confirming the observation in Table 4.

4.6 Human Evaluation

To demonstrate the effectiveness of the automatic evaluation metrics, we conduct human evaluations for summarization and perspective extraction. For summarization, we use criteria 1-5 from the GPT-4 evaluation prompt in Figure 3, excluding the last two related to formats. Two annotators are instructed to assign scores from 0 to 2 for each requirement, with 0 indicating complete dissatisfaction and 2 indicating full satisfaction. This approach yields a total score ranging from 0 to 10 for the generated summaries. We randomly select 50 samples from Perspectrumx and evaluate the summaries generated by GritLM@20 – Claude-3-Sonnet. Figure 7 displays the human annotation and GPT-4 scores. Human scores tend to be lower than GPT-4 scores, suggesting that GPT-4 might

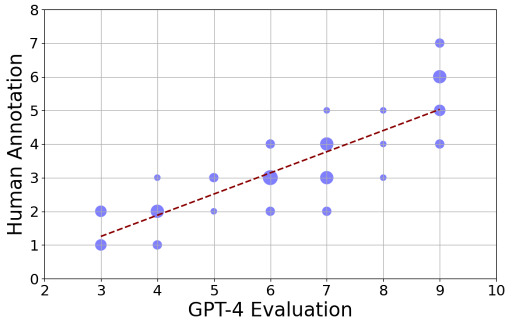


Figure 7: The relations between the GPT-4 scores and human scores for the multi-faceted summarization.

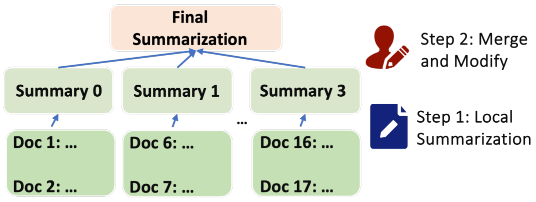


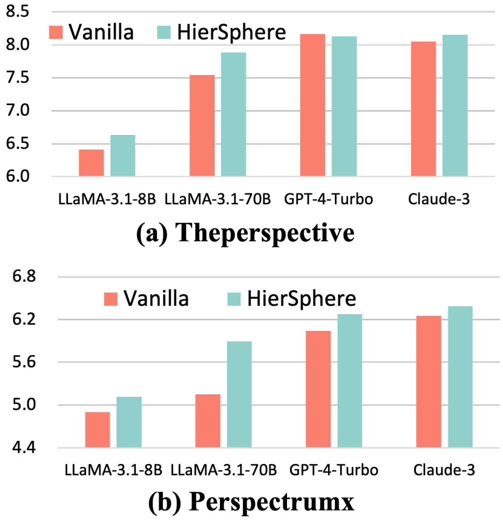
Figure 8: HierSphere for multi-faceted summarization.

overestimate performance compared to human assessments. However, there is a positive correlation between them, with a Pearson correlation coefficient of 0.70, indicating that GPT-4 scores could effectively reflect the summarization performance.

We also perform human annotation for perspective extraction. Two annotators determine whether the extracted perspectives are 'Entailment' or 'Not Entailment' to the gold standard. We randomly select 100 samples of generated perspectives from Perspectrumx using Claude-3-Sonnet for evaluation. The consistency between GPT-4 and human evaluation is 83% and 86%, respectively, demonstrating that GPT-4 is effective in evaluating the entailment of the extracted perspectives.

5 HierSphere

Based on the analysis, we find that the models fall short in the long context and perspective extraction. To further analyze whether we could enhance the model performance by reducing the context length, we propose a simple method - HierSphere - a hierarchical multi-agent summarization system (Figure 8). Specifically, after retrieving the relevant documents, we first adopt multiple agents to carry out local summarization on different sets of the retrieved documents. Then we adopt an editorial agent to further merge and modify the local summaries. In particular, the editorial agent is instructed to merge the perspective with equal semantic meaning and refine the perspectives given a one-sentence perspective demonstration. The prompt for the editorial agent is shown in Appendix C.



(a) Theperspective

(b) Perspectrumx

Figure 9: The performance of the summarization in vanilla and HierSphere using LLaMA-3.1-8B and LLaMA-3.1-70B on the Perspectrumx. The number of retrieval documents is 20 with GritLM.

The results of vanilla and HierSphere are shown in Figure 9, where we set four local agents and 20 retrieval documents. As we observe, the performance of HierSphere surpasses the vanilla in both Theperspective and Perspectrumx. For example, LLaMA-3.1-70B achieves a GPT-4 score of 5.89 in Perspectrumx, which outperforms vanilla by 0.78. The improvement can also be observed in closed-source models such as GPT-4-Turbo and Claude-3-Sonnet. It indicates that HierSphere could mitigate the challenges of the long context and perspective extraction and enhance the model performance on multi-faceted summarization.

6 Conclusion

We proposed a new task for multi-faceted perspective retrieval and summarization, which addresses the limitation of the existing work (Chen et al., 2019) that requires perspective pools in advance or overlooks the comprehensiveness, by adopting an RAG pipeline. For evaluation, we construct a dataset named PerSphere, and create a set of specific evaluation metrics tailored to this task. Performance of various LLMs on PerShpere highlighted the significant challenges inherent in such a complex task. Motivated by a thorough analysis, we devised a straightforward yet powerful multi-agent summarization method, HierSphere, designed to enhance model performance on PerSphere. The effectiveness of this method has been confirmed through experiments.

Limitations

In the task of argument mining (Ajjour et al., 2019; Friedman et al., 2021; Kamalloo et al., 2024), the volume of data is roughly comparable to ours. In real-world contexts involving large data sets, we could start with a coarse retrieval using BM25 on smaller subsets (like 10,000 or 5,000 documents), and then proceed to re-rank them. For our study, since the goal is to evaluate the comprehensiveness of retrieval models and the summarization abilities of LLMs, our dataset is adequate without needing to perform coarse retrieval from millions of documents. But we could further enhance the size of the data by adding irrelevant documents such as cnn_dailymail (See et al., 2017). In terms of retrieval, although it poses a challenge for current models, addressing this issue is not the focus of our study and could be considered in future work.

Ethical Consideration

We honor the ACL Code of Ethics. We collected data for Theperspective by automatically extracting data from www.theperspective.com. The CEO of the website, Daniel Ravner, granted us permission to extract and use their data for academic research. All annotators have received labor fees corresponding to the amount of their annotated instances.

Acknowledgement

This work has been supported by the National Key R&D program of China (grant No. 2022YFE0204900) and the National Natural Science Foundation of China Key Program under Grant Number 62336006. We thank Yulong Chen for helping us run experiments with closed-sourced commercial LLMs, and thank the ACL reviewers for their careful comments and suggestions.

References

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. LitSearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15068–15083, Miami, Florida, USA. Association for Computational Linguistics.

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2922–2932, Hong Kong, China. 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.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports*, 6(1):37825.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, et al. 2020. Corpus wide argument mining—a working solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7683–7691.

Mohamed Elaraby and Diane Litman. 2022. Arglegalsumm: Improving abstractive summarization of legal documents with argument mining. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Mohamed Elaraby, Yang Zhong, and Diane Litman. 2023. Towards argument-aware abstractive summarization of long legal opinions with summary reranking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7601–7612.

Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, et al. 2022. Growing polarization around climate change on social media. *Nature Climate Change*, 12(12):1114–1121.

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021.

Overview of the 2021 key point analysis shared task. In *Proceedings of the 8th Workshop on Argument Mining*, pages 154–164, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. 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.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (long papers)*, volume 1.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *arXiv preprint arXiv:2405.14486*.

Jie Huang and Kevin Chang. 2024. Citation: A key to building responsible and accountable large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 464–473.

Ehsan Kamalloo, Nandan Thakur, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, and Jimmy Lin. 2024. Resources for brewing beir: Reproducible reference models and statistical analyses. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1431–1440, New York, NY, USA. Association for Computing Machinery.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hao Li, Yuping Wu, Viktor Schlegel, Riza Batista-Navarro, Tharindu Madusanka, Iqra Zahid, Jiayan Zeng, Xiaochi Wang, Xinran He, Yizhi Li, and Goran Nenadic. 2024. Which side are you on? a multi-task dataset for end-to-end argument summarisation and evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 133–150, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yingjie Li and Cornelia Caragea. 2023. Distilling calibrated knowledge for stance detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6316–6329, Toronto, Canada. Association for Computational Linguistics.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085, Toronto, Canada. Association for Computational Linguistics.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Kuan-Chieh Lo, Shih-Chieh Dai, Aiping Xiong, Jing Jiang, and Lun-Wei Ku. 2021. Escape from an echo chamber. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 713–716, New York, NY, USA. Association for Computing Machinery.

Yun Luo, Zihan Liu, Yuefeng Shi, Stan Z. Li, and Yue Zhang. 2022. Exploiting sentiment and common sense for zero-shot stance detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7112–7123, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023a. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Yun Luo, Zhen Yang, Fandong Meng, Yingjie Li, Jie Zhou, and Yue Zhang. 2023b. Enhancing argument structure extraction with efficient leverage of contextual information. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7563–7571, Singapore. Association for Computational Linguistics.

Niklas Muennighoff, SU Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *ICLR 2024 Workshop: How Far Are We From AGI*.

C Thi Nguyen. 2020. Echo chambers and epistemic bubbles. *Episteme*, 17(2):141–161.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. 2023. Like-minded sources on facebook are prevalent but not polarizing. *Nature*, 620(7972):137–144.

Jan Heinrich Reimer, Alexander Bondarenko, Maik Fröbe, and Matthias Hagen. 2023. Stance-aware re-ranking for non-factual comparative queries. In *Proceedings of the 10th Workshop on Argument Mining*, pages 45–51, Singapore. Association for Computational Linguistics.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Allen Roush and Arvind Balaji. 2020. DebateSum: A large-scale argument mining and summarization dataset. In *Proceedings of the 7th Workshop on Argument Mining*, pages 1–7, Online. Association for Computational Linguistics.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. *arXiv preprint arXiv:2401.18059*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018a. ArgumenText: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 21–25, New Orleans, Louisiana. Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018b. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Shahbaz Syed, Timon Ziegenbein, Philipp Heinisch, Henning Wachsmuth, and Martin Potthast. 2023. Frame-oriented summarization of argumentative discussions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 114–129, Prague, Czechia. Association for Computational Linguistics.

Uğur Turan, Yahya Fidan, and Canan Yıldıran. 2019. Critical thinking as a qualified decision-making tool. *Journal of History, Culture & Art Research/Tarih Kültür ve Sanat Arastirmalari Dergisi*, 8(4).

Sander Van Der Linden. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature medicine*, 28(3):460–467.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

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

Zifeng Wang, Lang Cao, Benjamin Danek, Yichi Zhang, Qiao Jin, Zhiyong Lu, and Jimeng Sun. 2024. Accelerating clinical evidence synthesis with large language models. *arXiv preprint arXiv:2406.17755*.

Jianhao Yan, Yun Luo, and Yue Zhang. 2024. RefuteBench: Evaluating refuting instruction-following for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13775–13791, Bangkok, Thailand. Association for Computational Linguistics.

Jianhao Yan, Yun Luo, and Yue Zhang. 2025. Refutebench 2.0–agentic benchmark for dynamic evaluation of llm responses to refutation instruction. *arXiv preprint arXiv:2502.18308*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Processing Theperspective

During the dataset construction, we utilized LLMs solely to complete perspectives that are phrasal but not fully formed claims. This task is relatively straightforward, and the LLMs performed it with satisfactory results. Before implementing this procedure, we manually reviewed 25 samples and confirmed that the performance is satisfactory. In contrast, the Perspectrumx dataset does not involve LLM completion, and the task settings are significantly different, rendering direct comparisons invalid. We use GPT-4 to complete the incomplete perspectives in Theperspective.

Complete perspectives

Take a topic, a claim and an analytical perspective as inputs. The claim are the analysing stance to the topic and the perspective support the claim. The analytical perspective may not be a complete sentence or fully formed idea. If the given perspective is not a complete sentence, rewrite it into a clear, complete statement. If it is a complete sentence, just repeat it.
Directly output the sentence without any explanation.
Topic: {topic}
Claim: {claim}
Perspective: {perspective}
Answer:

To further enhance document diversity, we incorporate selected documents from the Perspectrum dataset into the Theperspective corpus. Specifically, we utilize NLTK (Natural Language Toolkit) to extract named entities (E) from the Theperspective documents. NLTK is a leading platform for building Python programs to work with human language data, providing easy-to-use interfaces to over 50 corpora and lexical resources, along with a suite of text-processing libraries. We use NTLK[11] to identify and extract entities from the Theperspective documents. And add the documents from Perspectrum with no entities existing Theperspective.

B Perspective Extraction

The prompts for extracting the perspective from documents and evaluation with golden perspectives are shown as follows:

Prompt for perspective extraction

Below is a claim and a document. Please summarize the document into a one-sentence perspective to support the claim. In your response, output the perspective surrounded by the key <p> </p>.
Claim: {claim}
Document: {document}

Prompt to evaluate perspectives

Determine whether the sentence 1 entails sentence 2.
Sentence 1: {sentence1}
Sentence 2: {sentence2}
Your response should be only a single word in [‘entailment’, ‘not_entailment’]

C Summarization Merge

We show the prompt to merge the local summarization and modify its perspectives.

Prompt for merging summaries

Given the summarization, merge them into one summairzation. Donot use your own knowledge.
The summarization should follow the requirements:
1. Each summarization should include both positive and negative claims, and only two.
2. Similar perspectives should be merged together with their references.
3. The reference of each perspective may exceed one.
4. Each perspective should be concise, arguing the claim from a specific aspect.
For example, ‘The law should not condone illicit behaviour’ or ‘Legalised prostitution still victimises the vulnerable.’
5. The output should be in a XML file. For example:
{document}
Summary: {summaries}

D Sub-Tasks

Apart from the pipeline solution, we analyze the possible sub-tasks to analyze model performance in detail, further analyzing the weakness of LLMs. We design specific prompts to evaluate LLMs as

[11] https://www.nltk.org/

Table 5: The results of the perspective extraction. GPT-4 refers to the entailment percent of the generated perspectives to the golden perspectives by querying GPT-4-Turbo, and BERT refers to BERTScore.

	Thepersective		**Perspectrumx**	
	BERT	GPT-4	BERT	GPT-4
LLaMA-3.1-8B	60.31	74.52	57.28	55.20
LLaMA-3.1-70B	**60.76**	80.24	57.55	60.02
GPT-4-Turbo	60.63	79.87	57.65	60.26
Claude-3-Sonnet	60.75	**80.24**	**58.11**	**60.88**

Table 6: The results (Macro-F1 score) of the sub-task stance detection using different LLMs.

	Thepersective	**Perspectrumx**
LLaMA-3.1-8B	86.15	61.62
LLaMA-3.1-70B	94.74	90.16
GPT-4-Turbo	94.56	88.80
Claude-3-Sonnet	91.19	88.75

shown in Table 9 and the prompt of perspective extraction is the same as Appendix B.

Perspective Extraction Given the document $d_m \in D^i$ and the corresponding claim c_0^i or c_1^i, we aim to extract a one-sentence perspective p^i of the document to support the claim. We directly adopt all the <document, claim> pairs for evaluation.

Stance Detection Given the perspective p^i and the claims c_0^i and c_1^i, we expect to determine which claim the perspective p^i supports. We adopt all the perspectives with the corresponding claims in a query for evaluation.

Reference Determination Given a perspective p and a document d, we expect to determine whether the document can serve as a reference to the perspective. To reduce the computation cost in this task, we assign each perspective p with a supporting document and a not-supporting document in the same query as the test set.

In the sub-task of perspective extraction, we evaluate the model performance using the golden documents, adopting both the entailment percent of GPT-4 and BERTScore (Zhang et al.)[12]. The results are shown in Table 5. We observe that the performance of different LLMs does not vary significantly, particularly in BERTScore. In GPT-4, the performance is relatively higher in Theperspective, but still significantly weak in Perspectrumx. The lower performance indicates that all models struggle to extract one-sentence perspectives, confirming the observation in Table 4.

Table 7: The results (Macro-F1 score) of the sub-task reference determination using different LLMs.

	Thepersective	**Perspectrumx**
LLaMA-3.1-8B	75.55	76.20
LLaMA-3.1-70B	82.20	79.78
GPT-4-Turbo	82.54	82.86
Claude-3-Sonnet	78.02	79.13

Table 8: The performance of the retrieving based on different retrievers with 100 documents. Recall refers to the ratio of relevant documents to the golden documents. Cover refers to the coverage values of the retrieved documents to the perspectives. In Theperspective, Recall equals Cover.

	Theperspective	**Perspectrumx**	
Metrics	Recall	Recall	Cover
BM25	90.25	76.67	76.64
E5-large	88.83	80.15	81.90
GTR-large	94.77	84.27	83.36
Ada-002	95.13	85.44	83.31
GritLM	99.55	86.99	85.61

We all adopt the macro-F1 score for evaluating these sub-task performances following (Glandt et al., 2021; Luo et al., 2022; Li et al., 2023). In the sub-task of Stance Detection, the results are shown in Table 6. Except for LLaMA-3.1-8B, other LLMs all achieve a significant performance, all above 85% F1 score. It indicates that the task of stance detection from perspective to claims is relatively simple to current well-performed LLMs.

In the sub-task of Reference Determination, the results are shown in Table 7. We observe that the model GPT-4-Turbo achieves the most significant performance to determine the reference with an F1 score of 82.54% in Theperspective and 82.86% in Perspectrumx. But Claude-3-Sonnet performs weakly in such a task and tends to output <Reference> labels which indicates a relatively weak capacity to distinguish the perspectives precisely.

E Retrieval Methods

BM25 (Robertson and Zaragoza, 2009): is a widely-used probabilistic ranking function for retrieval that scores documents based on the query terms they contain, considering term frequency, inverse document frequency, and document length normalization.

E5 (Wang et al., 2022): is pre-trained in a contrastive manner with weak supervision signals using human-curated text pair dataset for text embedding, whose embedding size is 1024.

[12] Specifically, we adopt the model deberta-v3 in the study.

Table 9: The specific prompts for sub-tasks which could also serve as the task definition. {·} denotes the input content.

Stance detection	Below is a perspective and two claims. Please determine which claim is supported by the perspective, Claim 1 or Claim 2. In your response, only output <Claim 1> or <Claim 2>.
	Claim 1: {claim_1}
	Claim 2: {claim_2}
	Perspective: {perspective}

Reference Determination	Below is a perspective and a document. Please determine whether the document can serve as an reference to the perspective. In your response, only output <Reference> or <Not_Reference>.
	Perspective: {reference}
	Document: {document}

Table 10: The specific prompts for sub-tasks which could also serve as the task definition. {·} denotes the input content.

Dataset	**Retrieval-Based**	**Extractive**	**Abstractive**	**Nonoverlapping**	**Comprehensive**	**# Domain**
Stab et al. (2018a)	✓	×	×	×	×	Various
Chen et al. (2019)	✓	×	✓	✓	×	Various
Roush and Balaji (2020)	×	✓	×	×	×	1 - Debate
Friedman et al. (2021)	×	×	✓	×	×	Various
Elaraby and Litman (2022)	×	✓	×	×	×	1 - Legal
Elaraby et al. (2023)	×	✓	×	×	×	1 - Legal
Reimer et al. (2023)	✓	×	×	×	×	Various
Li et al. (2024)	✓	×	✓	✓	×	1 - Debate
Our study	✓	✓	✓	✓	✓	Various

GTR-T5 (Ni et al., 2022): is a sentence-transformer model, which maps sentences and paragraphs to 768-dimensional dense vectors for the task of semantic search.

text-embedding-ada-002 [13]: is a closed-source text embedding model released by OpenAI.

GritLM (Muennighoff et al.): integrates text embedding and generation capabilities within a single language model, trained using instructional fine-tuning techniques.

F Retrieving with 100 Documents

The results are presented in Table 8. It is worth noting that these experiments are considerably more challenging in Perspectrumx compared to Theperspective. The model GritLM still performs the most satisfactorily, and the recall@100 results can be utilized for re-ranking, which could be future work.

G Data instances

We also show some data instances for our faceted perspective retrieval and summarization task as shown in Table 11. The first two instances are in Theperspective and the final one is in Perspectrumx. Figure 1 can be directly found in the link, which serves as an example of how we construct our data in Theperspective.

H Example for processing Perspectrum

Figure 10 shows how to process the data in Perspectrum to our data in Perspectrumx.

[13] https://platform.openai.com/docs/guides /embeddings/embedding-models

Table 11: Data instances in PerSphere.

Query	Claims	Perspectives	Reference
Surrealist Memes: Regression or Progression?	Surrealist memes represent progression in meme art.	Surreal memes are derived from a varied line of modern art traditions.	Doc 205
		Surrealist memes represent progression in meme art because they demonstrate that no message is necessary for impactful art.	Doc 364
	Surreal art is taking meme art back centuries (well, ok, decade)	"Surreal memes are internet elitism at its worst.	Doc 1138
		The perspective that surreal art is taking meme art back centuries is supported by the idea that much of its nuance is "Lost in Translation.".	Doc 858
Are Government Soda Taxes Fair?	The Soda Tax Is a Good Idea	The soda tax is a good idea because it can help curb obesity.	Doc 660
		The implementation of taxes on cigarettes successfully reduced consumption, suggesting a similar approach with soda could be effective.	Doc 345
		Investing in America supports the claim that the Soda Tax is a good idea.	Doc 550
	The Soda Tax Is Wrong	The Soda Tax is wrong because it infringes on free market principles.	Doc 769
		The soda tax is wrong because it only treats a symptom of a larger problem.	Doc 96
		The inconsistency of food stamp policies, which allow the purchase of sugary drinks, undermines the fairness of the soda tax.	Doc 523
Vaccination must be made compulsory?	We support the claim Vaccination must be made compulsory.	It is the state's duty to protect its community.	Doc 3692
		Compulsory vaccines are a financial relief on the health system.	Doc 8050
		Duty to protect the child.	Doc 3693, 8047, 2253, 5682, 6052
	We undermine the claim Vaccination must be made compulsory.	Compulsory vaccination violates the individuals' right to bodily integrity.	Doc 3695, 8052
		It is a parental right to decide about vaccinations for a child.	Doc 3696, 3695, 8047
		Vaccines have severe side effects.	Doc 3697, 3469

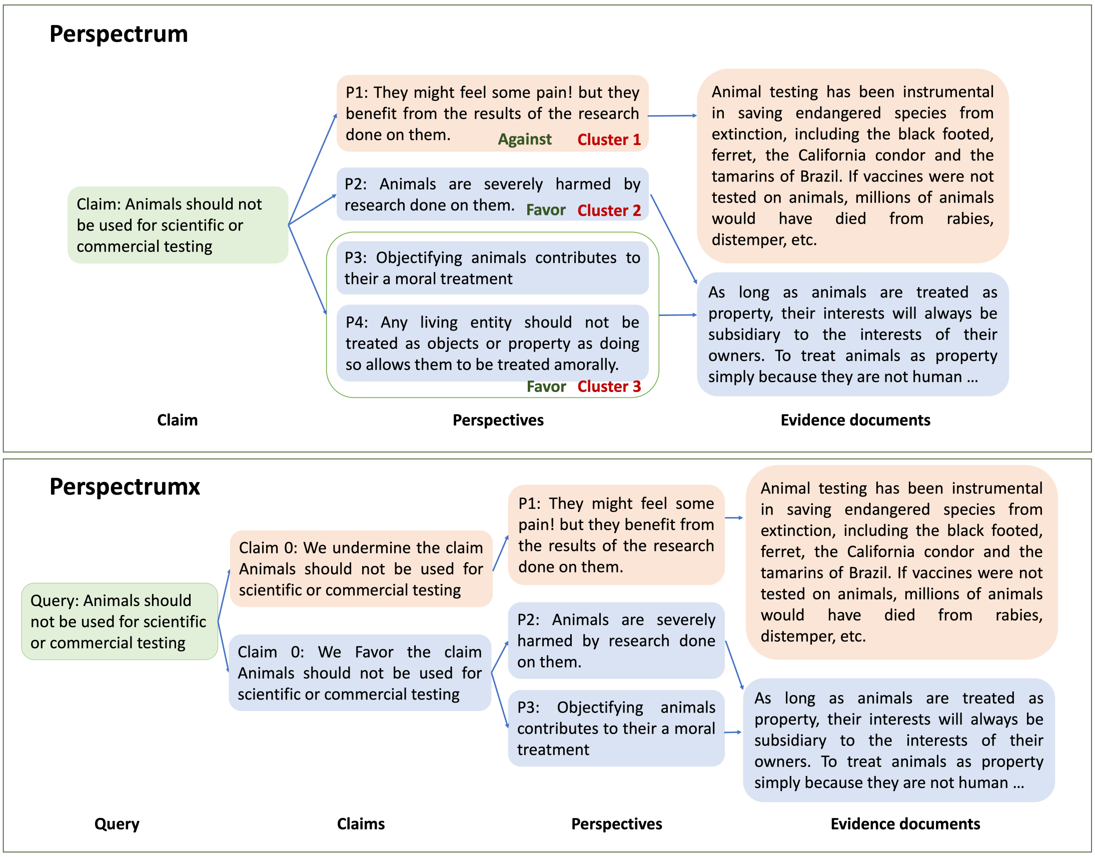

Perspectrum
Claim: Animals should not be used for scientific or commercial testing
P1: They might feel some pain! but they benefit from the results of the research done on them.
Against
Cluster 1
P2: Animals are severely harmed by research done on them.
Favor
Cluster 2
P3: Objectifying animals contributes to their a moral treatment
P4: Any living entity should not be treated as objects or property as doing so allows them to be treated amorally.
Favor
Cluster 3
Animal testing has been instrumental in saving endangered species from extinction, including the black footed, ferret, the California condor and the tamarins of Brazil. If vaccines were not tested on animals, millions of animals would have died from rabies, distemper, etc.
As long as animals are treated as property, their interests will always be subsidiary to the interests of their owners. To treat animals as property simply because they are not human ...
Claim
Perspectives
Evidence documents
Perspectrumx
Query: Animals should not be used for scientific or commercial testing
Claim 0: We undermine the claim Animals should not be used for scientific or commercial testing
Claim 0: We Favor the claim Animals should not be used for scientific or commercial testing
P1: They might feel some pain! but they benefit from the results of the research done on them.
P2: Animals are severely harmed by research done on them.
P3: Objectifying animals contributes to their a moral treatment
Animal testing has been instrumental in saving endangered species from extinction, including the black footed, ferret, the California condor and the tamarins of Brazil. If vaccines were not tested on animals, millions of animals would have died from rabies, distemper, etc.
As long as animals are treated as property, their interests will always be subsidiary to the interests of their owners. To treat animals as property simply because they are not human ...
Query
Claims
Perspectives
Evidence documents

Figure 10: An example of data processing from Perspectrum to Perspectrumx.