Web-Augmented Multi-Perspective RAG Summarization

1 Introduction

Many people rely on LLMs and other generative text tools to facilitate text-related tasks. One widely used task is text summarization, as users would want long texts (like papers) to be summarized into a form that is shorter, and easier to understand. While LLMs are able to read text at high speeds and with generally highly accurate understanding, another key aspect to consider is how well they balance varying opinions on a topical summary? It is essential for text summaries to not only condense key ideas, but also integrate varying claims with distinct perspectives to produce more well-balanced results.

As such, our primary research question is as follows:

Central Research Question: How does using web-search grounding help improve the overall quality of multi-perspective summarization compared with RAG pipeline that relies solely on static corpora?

1.1 Design Rationale

We selected the PerSphere dataset (Luo et al. (2025b)) and used it as a foundation for the following reasons:

1. **Intrinsic balance and annotation quality:** Each query is a controversial question which already contains gold, stance-balanced oppositional claims, supporting perspectives and evidence manually curated by annotators, providing a strong baseline ground truth to evaluate the true impact of web-grounding on retrieval and multi-perspective-summarization quality.

2. **Controlled comparability:** Using the same PerSphere baseline provides direct, one-to-one comparison between the static and web-augmented pipeline for offline retrieval metrics and overall summarization quality.

3. **Relevance to multi-perspective summarization:** The broad coverage of controversial topics (education, politics, health, etc.) alongside validated oppositional claims for each query helps work with a defined set of queries which can address multi-perspective summarization. In contrast, alternative datasets or varied pipelines can introduce loosely defined queries with ambiguous or overlapping perspectives, which can complicate evaluation.

1.2 What is multi-perspective summarization?

Multi-perspective summarization builds atop the PerSphere framework from Luo et al. (2025b) which includes data from two primary sources: ThePerspective.com and Perspectrum. From ThePerspective.com they obtain editorial articles around controversial topics each with a central query (eg: Should phones be banned in schools?) and two opposing claims that respond to the query (eg: 1. Yes, phones should be banned 2. No, phones should not be banned). Additionally, each article includes several perspectives supporting each claim (eg: 1. banning phones promotes student focus. 2. phones help students stay in contact with parents.). The corresponding set of paragraphs associated with each perspective is marked as the associated gold evidence document. Of particular note, each perspective has only one evidence document for ThePerspective.com whereas Perspectrum expands this framework to include multiple documents that support perspectives via data collected from online debate forums.

Together, these sources form the PerSphere dataset which provides gold-standard entries as such:

{query, two opposing claims, perspectives, evidence documents}

This the foundation of our experimental work as it provides a strong framework to extend with web-search grounding and analysis.

Building atop this foundation, we extend Luo et al. (2025b) by introducing a dynamic web-retrieval layer that enables the discovery of both corroborative and novel perspectives from live web data. Unlike static corpora, web results can surface arguments that either align with existing perspectives (e.g., reinforcing known claims with fresher evidence) or introduce new yet valid perspectives that were not covered in the original dataset.

Therefore, our framework first determines whether each retrieved web document semantically correlates with an existing perspective. If no strong correlation is detected, the model tests whether it constitutes a new perspective candidate - a concise, stance-bearing statement entailed by the document and relevant to one of the opposing central claims associated with the query. Entailment-validated new perspectives are appended to the perspective set and subsequently evaluated for coverage and diversity in our extended metrics.

First, we define a perspective as a concise supporting statement that expresses a binary stance (either for or against) one of the two opposing claims derived from the query in ThePerspective.com dataset. Each perspective is ideally non-overlapping with others and is entailed by one or more evidence documents. This definition extends Luo et al. (2025b) but is adapted for our web-augmented pipeline, where new web documents may introduce novel yet stance-consistent perspectives.

Then, multi-perspective summarization refers to the process of generating semi-structured summaries that cover non-overlapping perspectives which support opposing claims on a given query, while maintaining factual accuracy with respect to source documents. For example, given the query ”Should we ban phones from schools?”, the model should summarize both opposing claims:

(1) ”Schools should be banning phones” with perspectives like ”to emphasize academic focus” and ”to promote in-person socialization” [Cite opposing sources]

(2) ”Schools should allow phones” with perspectives of ”allowing emergency communication” and ”help develop self-control”. [Cite supporting sources]

We believe that by augmenting the multi-perspective summarization pipeline from Luo et al. (2025b) with web-search grounding can promote more recent, more balanced and higher quality multi-perspective summaries.

However, since web-augmentation dynamically modifies the original corpus the given offline metrics (like Recall@k and Cover@k) are no longer sufficient for measuring the retrieval effectiveness in our setting.

As such, we extend Luo et al. (2025b) by introducing novel web-grounded retrieval metrics to quantify web-augmented retrieval coverage, perspective and retrieval balance in our dynamic document context.

Specific formulations of these metrics and their evaluation procedures are detailed in the Approach section below.

2 Related work

Our main research inspiration is based on Luo et al. (2025b) which proposed the PerSphere benchmark for multi-perspective retrieval and summarization in NLP. As per Luo et al. (2025b), prior works in retrieval augmented generation (RAG) summarization systems tend to focus on relevance of retrieved documents without consideration for multi-perspective document retrieval. This tends to produce summaries that are one-sided which serve to promote online echo chambers and biases.

Luo et al. (2025b) addresses this by introducing the Persphere framework. Firstly, the paper highlights that Persphere comprises of two datasets — one taken from ThePerspective.com (with permission) and one created from the Perspectrum dataset, totaling over 1000 examples. Secondly, a benchmark that measures the perspective balance of retrieved documents in terms of Recall@k and Cover@k is used in the paper. Once retrieval is measured, they used GPT-4 is used to rate the final summary, on a scale of 1 - 10, based on key metrics like the presence of varying perspectives, lack of overlap between perspectives, and such. They found the best retrieval (GritLM) and model (GPT-4-Turbo) achieved high recall@k of around 70% with a fairly average final GPT-4 score of 6/10 on the harder Perspectrumx subset of the data.

Luo et al. (2025b) address a critical gap in comprehensive RAG LLMs which can serve to produce more balanced and fair summaries, especially in controversial domains like Politics. On the other hand, Luo et al. (2025b), operates on a fairly limited dataset (about 1,000 claims with around 12,000 documents) without web-based grounding; thus providing a fantastic opportunity to extend their framework with simple web-search tooling to measure the impact that web-grounding can have on retrieval quality and final comprehensiveness rankings on the Persphere benchmark.

Secondly, Liu et al. (2024) documents the lost in the middle challenge with LLMs using long-context inputs (roughly greater than 2,048 tokens). In particular, they found that when using transformer based models(both decoder-only and encoder-decoder), a robust U-shaped performance curve emerges where performance is best (at least 56% accuracy) when the relevant information is present at the beginning (position 1 - 5) or ending (position 16 - 20) of the input sequence, and accuracy dips consistently relevant information is embedded in the middle of the input even with models designed to handle tens of thousands of input tokens.

Liu et al. (2024) observes a critical, cross-architecture (within transformer family) performance bottleneck of LLM usage for summarization where the position of relevant information significantly affects the performance of LLMs in summarization tasks. One weakness of the paper is that non-transformer architectures like state space models, like Mamba, are not analyzed and this trend may not generalize beyond transformer-based language models. In relevance to our main goal, this paper serves as a key reference on where to positionally encode web-derived information in our multi-perspective summarization framework to maximize the downstream comprehensiveness.

Thirdly, Ding et al. (2025) introduces the latest large-scale, multi-turn, multi-party stance detection. This improves upon prior work which typically targets small number of fixed targets (usually 5 - 50). They collect around 17,000 samples from Weibo with 280 targets and 6 areas. The samples are annotated with 3 options: Either in favor of a target, against or neutral to with a strong 0.83 Cohen's Kappa for human annotator agreement.

The major technical contribution of Ding et al. (2025) is the SITPCL (Speaker Interaction and Target aware Prototypical Contrastive Learning) model which includes a speaker interaction encoder to model the relationship between and within entities, as well as a target aware contrastive learning to increase the representation of unseen target stance. This model outperformed various baselines across Llama-3 and Roberta, though only achieving a fairly poor F1 of 0.40 to 0.50.

Thus, while Liu et al. (2024) observes a new SOTA performance on many-target-stance-detection, it's still lackluster performative F1 score indicates the practical challenges of stance detection in unseen conversational contexts. For instance, they observed that comments which employ sarcasm have a more complex reasoning that confuses these stance detectors in around 21% of cases. One advantage of this paper, however, is that they expand stance-detection to large-scale targets which is an important contribution as single documents can easily contain multiple targets and being able to effectively detect stances on multiple targets within text is critically important. In relevance to our main goal, this paper serves as a key reference on the current SOTA challenges in multi-target stance detection and how we should carefully approach and limit the scale of stance detection on web-source data to properly to maximize the downstream comprehensiveness of our multi-perspective summarization framework.

In addition, Luo et al. (2025a) proposes a method to make recursive summarization more efficient and more query aware in RAG systems. Traditional summary trees often have a problem with large numbers of redundant nodes, which can cause an increase in computation time and negatively impact the answer of questions that require summarization. The paper introduces DTCRS as a solution to this problem, first determining whether a summary tree is necessary by analyzing the question, then looking at the question's embeddings and setting them as the starting cluster points. Through recursive clustering and text summarization, DTCRS successfully creates the hierarchical summary tree that can be used to output a simple summary of a question asked about documents such as the example in their sixth figure. The results show that the method performs better on multiple benchmarks. Limitations include sometimes making wrong classifications or if the document is too long, longer than the model's input limit, then the tree could be incomplete and not summarized effectively. In relevance to our project, recursive tree based summarization could be implemented to make the summarization process more efficient and tailored to output based on the given query or input for the LLM.

Finally, when facing the challenges of summarizing multilingual or multiple articles on the same topic, Zhang et al. (2025) focuses research on creating summaries given a cluster of new articles about the same main event, with particular attention to multilingual and localized contexts. The paper proposes McMs, a framework for data sharpening designed to address issues in current methods that tend to prioritize sentences at the beginning of articles and suffer from heavy-tailed sentence distributions. In this framework, sentences are grouped into clusters based on their semantic similarity, allowing the model to better capture diverse perspectives within event coverage.

Zhang et al. (2025) further apply joint finetuning to the language model to enhance its ability to capture localized context across languages and sources. This approach improves the model's capacity to identify and integrate relevant details from multilingual inputs. The authors evaluate their method using the SeaSumm-v1 data set for the Clust-McMs pipeline task and GlobeSumm for the unified McMs task, measure peformance with ROGUE, Event Coverage (Eve-Cov), and Entity Faithfulness (Ent-Faith) metrics. However, because the localization experiments were based on the GlobalSumm dataset, which had extensive coverage of 26 languages, the localization tasks became quite complex, and having up-to-date data for all languages is a challenging component, so small-scale data will not yield meaningful results. This research can be relevant to our project if we come across anything that is not English, since not everything online is only in English.

3 Approach

Our main contribution extends the work of Luo et al. (2025b) by incorporating web-based retrieval, and entailment and novelty validation, to increase the overall quality of multi-perspective summarization. For a high level view of our pipeline please consider Figure 1.

1. **Dataset Note:** We will primarily be focusing our experimental efforts on ThePerspective subset of PerSphere, from Luo et al. (2025b). If time permits, we may expand experiments to Perspectrumx or even other relevant datasets.

2. For each query in ThePerspective:

   (a) **Corpus Retrieval:** Apply TF-IDF over the static Perspective corpus to rank documents based on term frequency and inverse document frequency scores.

   (b) **Web Retrieval**: Using the same query we perform a live web-search to obtain the top-x most relevant results ($x \leq k$). We may adjust this to searching specific news websites or targeted sources if raw web search proves inconsistent for our pipeline.

   (c) **Entailment and Novelty Validation:** Each web document is evaluated by GPT-5-Nano through a two-stage process (with manual human validation as specified later for validation):

   - **Perspective Correlation:** Compare the new web document against all associated offline perspectives to determine it matches any of them. If yes, then it gets added to the evidence pool of said matched perspective.
   - **Novel Perspective Induction:** If no correlation is found between the web document and pre-existing perspectives then GPT-5-Nano proposes a concise novel perspective in alignment with either the pro or con claim and this gets added to the associated information with this query.

   This key step ensures that the pipeline can both integrate novel, stance-consistent perspectives from the web and reinforce known perspectives.

   (d) **Hybrid Combination:** Merge the static corpus with the validated web corpus.

   (e) **Summarization:** Input the merged corpus as context into a generative LLM, such as Llama-3.1-8B-Instruct to produce a web-enhanced multi-perspective summary.

   (f) **Evaluation**: Measure retrieval and summarization quality using both baseline (static) and web-augmented (dynamic) metrics built from a foundation of Luo et al. (2025b).

(g) **Manual Human Validation:** Conduct human evaluation on a representative subset of ThePerspective to verify LLM-as-Judge reliability for summarization, perspective correlation and novel perspective induction.

In our extension, we introduce a novel web-retrieval and validation layer that aims to integrate both correlated and novel perspectives discovered from web content. By integrating both web-grounding and static retrieval we hope our pipeline can systematically measure how the balance and factuality of multi-perspective summaries can be increased as compared with the purely offline corpus-based work of Luo et al. (2025b).

3.1 Baseline Components

- **Retrieval:** TF-IDF - A simple baseline method which ranks documents by boosting the score based on increased term frequency and decreasing the score based on the inverse document frequency of query terms.

- **Summarization Model:** Llama-3.1-8B-Instruct - A generative instruction-tuned smaller-scale model which we chose for its balance of fluency and compute efficiency. It was also used by Luo et al. (2025b), so it allows comparability with their work on offline metrics. We may consider alternative variants (e.g., Llama-3.2-1B-Instruct or Phi-3-Mini) if compute limitations emerge, however, this is our planned model of choice for the moment.

- **Rankings / Entailment Model:** GPT-5-Nano - Will be used for both entailment classification for web documents, and as the automatic judge for multi-perspective summarization scores. These will be cross-validating with human annotations to establish effectiveness. And again, lighter models may be considered depending on compute limitations.

3.2 Evaluation Design

We will keep Cover@k and Recall@k as baselines for fair comparison with the prior work of Luo et al. (2025b). That being said, we are introducing web-extensions of these metrics to explicitly account for the non-static corpus of documents that will be utilized in our pipeline.

Retrieval Quality Metrics. Here let's define key variables that will be utilized for our web-augmented metrics:

- R_i - The set of retrieved documents (includes both web and offline unless otherwise specified)

- D_i - The gold documents for a query as annotated and provided by Luo et al. (2025b) for ThePerspective subset.

- W_i^{ent} - The set of relevant web documents as entailed from prior perspectives or novel.

- P^{gold} and $P^{\text{new,ent}}$ - Corresponding sets of perspectives from the offline corpus and web-retrieval.

Now we present our novel metrics for our web-augmented multi-perspective summarization pipeline:

- **Perspective Expansion Rate (PER):**

$$PER = \frac{|P^{new,ent}|}{|P^{gold}|}$$

  Will be used to measure the degree of novel and entailed perspectives that web-augmentation provides over the baseline gold standard of perspectives.

- **W-Rp@k:**

$$W\text{-}Rp@k = \frac{w_i : w_i \Rightarrow (P_j \text{ or } P_{new})}{|W_i|}$$

  This metric calculates the proportion of retrieved web documents entailed by either known or novel perspectives.

- **d-Recall@k:**

$$dRecall@k = \frac{|R_i \cap (W_i^{ent} \cup D_i)|}{k}$$

  For the recall variant, we adapt the Recall@k metric to quantify how many of the top retrieved documents(offline or web + entailed) are truly relevant in our hybrid pipeline.

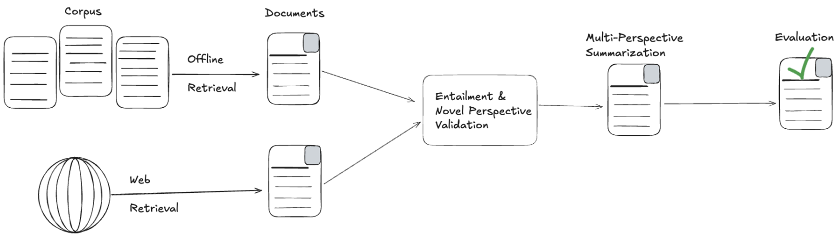

Figure 1: Extended pipeline for web-grounded multi-perspective summarization.

- **d-Cover@k:**

$$dCover@k = \frac{|P_i^{covered} \cup P_i^{new,ent}|}{|P_i^{gold}| + |P_i^{new,ent}|}$$

This measures perspective-level completeness, indicating what proportion of all relevant perspectives were properly harnessed

By using these metrics we aim to properly measure the impact of web-integration of key aspects of our web-augmented retrieval pipeline such as high-quality novel perspectives, the proportion of web documents that are entailed as high-quality, and the completeness of relevant web-enhanced documents and perspectives.

Multi-Perspective Summarization Quality Evaluation

We will rate each summary by GPT-5-Nano (LLM-as-Judge) via 5 key criteria extended from Luo et al. (2025b) with a specialized component for web-quality measurement at the end. Each component is rated on ascale from 0 to 2 where 0 means *completely unsatisfactory*, 1 means *neutral satisfactory*, and 2 means *fully satisfactory*.

1. Distinctness of perspectives which measures if perspectives overlap with each other or not.
2. Factual consistency between the perspective and associated evidence document.
3. The accuracy of the perspective which measures it's relevance to the core claim.
4. Completeness of offline perspective which measures how much of the pipeline summary includes content from the gold summary (which utilizes only the true annotated perspective + evidence pairs from ThePerspective subset). Here we will look to see how much of offline-gold perspectives from the gold summary are included, but we shall not penalize for novel relevant perspectives as it is a core feature of our pipeline.
5. Factuality of responses to make sure that the summary doesn't include information that is hallucinated or inaccurate with respect to provided evidence.

Human-Driven Manual Validation

1. From ThePerspective subset we will sample 30 queries.
2. Using both the static pipeline (as specified by Luo et al. (2025b)) and our web-augmented pipeline, we shall generate multi-perspective summaries for the aforementioned 30 queries.
3. Two of us can rate each summary using the five-criterion rubric (0 - 10 score)
4. Compute Pearson correlation between human and LLM-as-Judge scores to assess consistency.
5. Estimate inter-annotator agreement via Cohen's kappa.
6. For web retrieval each web document can be labeled as either novel and entailed or expanding a prior perspective with respect to either oppositional claim.
7. Compare these labels with the GPT-5-Nano classifications to compute entailment precision and validate *W-Rp@k*.

By combining both static and dynamic web-augmented metrics our experiment aims to measure the impact that web-augmented retrieval has on the overall factuality and balance of multi-perspective summarization.

4 Data

We will be primarily working with the PerSphere dataset and live web-documents. As described in Luo et al. (2025b), this dataset is composed of balanced queries and associated documents collected from two sources:

- The first subset (Theperspective) has data collected from the website THEPERSPECTIVE, which stores many editorial articles, each containing a query title with two controversial claims supported by multiple perspectives. Corresponding to each perspective is one and only one evidence document.
- The second subset (Perspectrumx) has data collected from the dataset Perspectrum, which stores claims from online debates, perspectives supporting claims, and evidence for each perspective. Unlike Theperspective, multiple evidence documents support each perspective. The subset created in PerSphere uses these claims as queries, then connects them to statements saying ”We favor/undermine the claim query”. A variety of documents from Perspectrumx are added to Theperspective so the latter can have a more diverse document collection.

5 Tools

We aim to use standard NLP tools including Python, PyTorch, and any requisite NLP libraries, such as HuggingFace, that will be necessary for model training. Additionally, we will look to use Google Collab for GPUs to train our models during the course of this experiment. If anything comes up during the data processing, training, or evaluation that requires a specific library or toolkit, it can be documented later.

References

Ding, Y., He, K., Li, B., Zheng, L., He, H., Li, F., Teng, C., and Ji, D. (2025). Zero-shot conversational stance detection: Dataset and approaches. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3221–3235, Vienna, Austria. Association for Computational Linguistics.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Luo, G., Jian, Z., Qiu, W., Wang, M., and Wu, Q. (2025a). DTCRS: Dynamic tree construction for recursive summarization. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10948–10963, Vienna, Austria. Association for Computational Linguistics.

Luo, Y., Li, Y., Hu, X., Qi, Q., Guo, F., Guo, Q., Zhang, Z., and Zhang, Y. (2025b). PerSphere: A comprehensive framework for multi-faceted perspective retrieval and summarization. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21790–21805, Vienna, Austria. Association for Computational Linguistics.

Zhang, L., Zou, B., and Aw, A. (2025). Enhancing event-centric news cluster summarization via data sharpening and localization insights. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16412–16426, Vienna, Austria. Association for Computational Linguistics.