Disco-RAG: Discourse-Aware Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) has emerged as an important means of enhancing the performance of large language models (LLMs) in knowledge-intensive tasks. However, most existing RAG strategies treat retrieved passages in a flat and unstructured way, which prevents the model from capturing structural cues and constrains its ability to synthesize knowledge from dispersed evidence across documents. To overcome these limitations, we propose Disco-RAG, a discourse-aware framework that explicitly injects discourse signals into the generation process. Our method constructs intra-chunk discourse trees to capture local hierarchies and builds inter-chunk rhetorical graphs to model cross-passage coherence. These structures are jointly integrated into a planning blueprint that conditions the generation. Experiments on question answering and long-document summarization benchmarks show the efficacy of our approach. Disco-RAG achieves state-of-the-art results on the benchmarks without fine-tuning. These findings underscore the important role of discourse structure in advancing RAG systems.[1]

1 Introduction

The advent of large language models (LLMs; Touvron et al. 2023; Yang et al. 2025; Achiam et al. 2023) has advanced research progress in natural language processing (NLP), achieving competitive performance across a wide range of tasks, including question answering (Wu et al., 2025a; Lee et al., 2025a; Zhang et al., 2025b), document summarization (Mondshine et al., 2025; Liu et al., 2025a; Wang et al., 2025a; Luo et al., 2025), and text generation (Duong et al., 2025; Bigelow et al., 2025; Que and Rong, 2025; Zhang et al., 2025a). However, due to the reliance on static training corpora, LLMs can be inadequate for knowledge-intensive scenarios, such as handling domain-specific knowledge, proprietary data, or information requiring real-time updates (Chang et al., 2025; Lee et al., 2025b; Yue et al., 2025; Wang et al., 2024b; Xia et al., 2025). Retrieval-Augmented Generation (RAG) has been proposed as a suitable strategy by integrating an external knowledge component through retrieval-based mechanisms (Lewis et al., 2020; Asai et al., 2024; Chan et al., 2024).

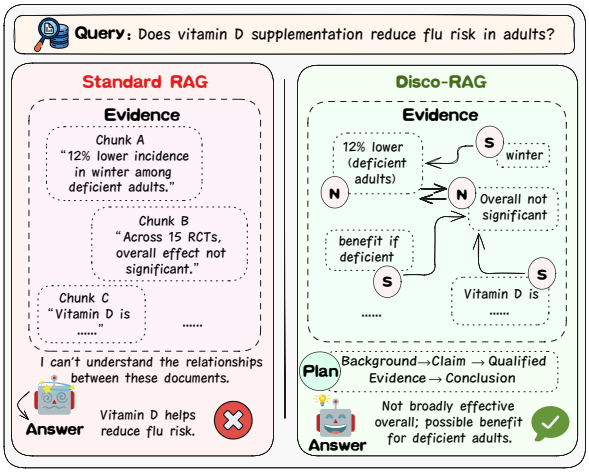

Figure 1: Comparison between standard RAG and Disco-RAG. While standard RAG retrieves isolated chunks without structural links, Disco-RAG organizes evidence into discourse structures (trees & graphs). Here, S denotes *Satellite* (the supplementary part), and N denotes *Nucleus* (the core part).

In standard RAG pipelines, external documents are segmented into chunks, which are then encoded into vectors and stored in a database. At query time, relevant chunks are retrieved to provide contextual grounding for the LLM (Lewis et al., 2020). One important but insufficiently addressed limitation of existing RAG systems concerns the mismatch between retrieval granularity and generative understanding. While retrieval modules return relevant chunks, these chunks are often fragmented in discourse, resembling scattered pieces of evidence without clear logical connections (Edge et al., 2024; Su et al., 2025). This manifests at two levels.

[1]Code is available at https://anonymous.4open.science/r/Discourse-RAG

First, *intra-chunk structural blindness*: within each chunk, RAG systems may fail to capture internal discourse. Second, *inter-chunk coherence gaps*: across multiple chunks, RAG systems may struggle to identify rhetorical connections. As depicted in Figure 1 (left), *Chunk A* mentions *a 12% lower incidence*, while *Chunk B* notes *no significant overall effect*. Without recognizing that the former is a conditional finding (*e.g.,* among deficient adults in winter), standard RAG tends to overgeneralize and incorrectly conclude that *vitamin D reduces flu risk*. These deficiencies prevent effective resolution of conflicting claims, as standard RAG approaches lack the capacity to organize retrieved evidence through a higher-level causal flow. This leaves the final LLM generator to grapple with a *bag of facts* rather than a coherent *line of reasoning*.

Recent investigations have revealed that integrating discourse knowledge into LLMs can improve downstream performance (Gautam et al., 2024; Liu and Demberg, 2024) and alleviate hallucinations (Liu et al., 2025b). These findings highlight the drawback of relying solely on flat sequential representations and underline the benefits of discourse for context engineering (Ma et al., 2025; Mei et al., 2025). Building on these insights, the present work aims to investigate whether explicitly modeling and providing discourse information to the LLM can improve generation quality in the context of RAG. To answer this, we propose `Disco-RAG`, a framework that constructs local discourse trees for each retrieved chunk and infers inter-chunk coherence relations across chunks to form a rhetorical graph. To synthesize information, rather than merely concatenating it, the text generator needs not only to understand the relations between evidence but also to strategize how to present them. This requires a high-level plan to orchestrate the narrative flow. We thus introduce a discourse-aware planning module that enables the model to dynamically generate a plan to guide the generation. As shown in Figure 1 (right), the discourse-aware process enables the model to infer that *vitamin D is not broadly effective but may benefit deficient adults under specific conditions*, producing more faithful answers and aligning with the underlying evidence.

In our experiments, we evaluate `Disco-RAG` on three benchmarks, `Loong` (Wang et al., 2024a), `ASQA` (Stelmakh et al., 2022), and `SciNews` (Liu et al., 2024). Consistent improvements are observed compared with standard RAG systems and state-of-the-art (SOTA) methods. On the Loong benchmark, our approach delivers gains of up to 10.0 points in LLM Score. On the ASQA dataset, our method exceeds the best existing systems on Exact Match and DR Score by notable margins. On the SciNews benchmark, `Disco-RAG` establishes new SOTA performance across most evaluation metrics.

In summary, our contributions are as follows:

- We present `Disco-RAG`, an inference-time strategy that explicitly injects discourse knowledge into the RAG pipeline to alleviate the discrepancy between chunk-level evidence and discourse-level reasoning.
- We propose a modeling method that combines intra-chunk discourse trees, inter-chunk rhetorical graphs, and discourse-driven plans to capture local hierarchies, cross-passage coherence, and argumentative flow.
- We conduct experiments on knowledge-intensive QA and summarization tasks, demonstrating consistent gains over strong RAG baselines. Analysis studies further confirm the efficacy of discourse-aware guidance in enhancing generation correctness, coherence, and factuality.

2 Related Work

2.1 Structure-Aware Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) enhances LLMs in knowledge-intensive tasks by retrieving external evidence (Lewis et al., 2020). However, conventional RAG methods typically treat retrieved chunks as isolated and flat sequences, overlooking their structural interconnections. To mitigate this, recent research has explored structure-aware variants of RAG. Graph-based methods (Nigatu et al., 2025; Hu et al., 2025; Wu et al., 2025b; Zhu et al., 2025) such as GraphRAG (Edge et al., 2024) and KG-RAG (Sanmartin, 2024) organize evidence into knowledge graphs, while subsequent work has improved retrieval by simulating human memory mechanisms (Gutierrez et al., 2024; Gutiérrez et al., 2025) or enriching graph semantics (Liang et al., 2025). Other approaches construct structured subgraphs for coherence (Mavromatis and Karypis, 2025; Li et al., 2025a), or employ alternative formats like hierarchical graphs (Zhang et al., 2024; Wang et al., 2025b; Huang et al., 2025), semantic chunking (Wang et al., 2025c; Qu et al., 2025; Zhao et al., 2025), trees (Fatehkia et al., 2024; Sarthi et al., 2024), and tables (Lin et al., 2025). More

adaptive strategies dynamically select structures based on context (Li et al., 2025b). Despite these advances, most efforts emphasize surface-level associations (*e.g.,* linking entities) while largely overlooking the rhetorical structure that governs causal flow, evidence presentation, and conclusion formulation. This hinders logical depth and discourse coherence, which our work seeks to address.

2.2 Rhetorical Structure Theory for Text Generation

Rhetorical Structure Theory (RST; (Mann and Thompson, 1987, 1988)) is a discourse framework that models hierarchical dependencies and rhetorical relations among Elementary Discourse Units (EDUs). It distinguishes between *nucleus* and *satellite* units, connected by relations such as *Elaboration*, *Causality*, and *Contrast*, forming tree structures that reflect communicative intent. Foundational work (Marcu, 1997, 1999; Mann and Thompson, 1987; Bhatia et al., 2015; Hayashi et al., 2016) has established strong correlations between rhetorical structure and human text planning (Adewoyin et al., 2022). Later studies have leveraged RST by converting trees into dependency graphs or imposing structural constraints to improve coherence and consistency in neural generation models (Chistova, 2023; Zeldes et al., 2025; Chistova, 2024; Maekawa et al., 2024). More recent efforts have integrated RST into LLMs to improve cross-sentence reasoning and enhance both structural integrity and interpretability of generated outputs (Liu et al., 2023; Liu and Demberg, 2024). Compared with shallow discourse markers or sentence-level connectives, the present work extends RST modeling to the RAG setting by explicitly encoding the deeper structure of retrieved passages and highlighting the importance of hierarchical discourse.

3 Proposed Method

Method Overview. We formalize the standard RAG as a conditional generation problem. Given a query q and a set of Top-k retrieved chunks $\mathcal{C}(q;\mathcal{D}) = \{c_1, c_2, \ldots, c_k\}$ from a corpus $\mathcal{D}$, the output is $y = \arg\max_{y'} P(y' \mid q, \mathcal{C}(q;\mathcal{D}))$, where $P(\cdot)$ denotes the conditional distribution of the generator. To overcome the limitations of the retrieval-and-concatenation paradigm, we propose `Disco-RAG` to augment standard RAG with rhetorical parsing and discourse-aware planning.

As illustrated in Figure 2, our pipeline consists of three main stages. (1) We delve into each chunk c_i to uncover its internal logical hierarchy by constructing an intra-chunk RST tree t_i, (2) We zoom out to map the relational landscape across all chunks $\mathcal{C}$ via an inter-chunk rhetorical graph $\mathcal{G}$, and (3) We apply a discourse-driven planning module that devises a blueprint $\mathcal{B}$ based on $\mathcal{T} = t_{i=1}^{k}$ and $\mathcal{G}$ to guide the final generation process.

We hypothesize that under identical retriever and decoding conditions, explicitly injecting discourse knowledge improves the correctness, coherence, and factual consistency of generated text. Here, rhetorical modeling serves as a *knowledge-level prior*, while planning offers *reasoning-level guidance*, jointly inducing stronger structural biases than standard RAG. The following paragraphs provide a detailed account of each component.

Intra-Chunk RST Tree. For each retrieved chunk c_i, we construct an RST tree t_i using an LLM-based RST parser $\mathcal{A}$ to model local coherence.[2] Given c_i, parser $\mathcal{A}$ jointly performs elementary discourse unit (EDU) segmentation and RST parsing, producing a sequence of EDUs $\{e_{i_1}, \ldots, e_{i_m}\}$, nucleus and satellite role assignments, and rhetorical relations among EDUs. Formally, $c_i \xrightarrow{\mathcal{A}} t_i = (V_i, E_i)$, where $V_i = \{e_{i_1}, \ldots, e_{i_m}\}$ is the set of EDU nodes, $\mathcal{R}$ is the set of rhetorical relations (*e.g.*, *Elaboration*, *Contrast*, and *Cause*), and $E_i \subseteq V_i \times V_i \times \mathcal{R}$ is the set of directed connections labeled with relation types. The symbol $\times$ denotes the *Cartesian product*. The top-middle panel of Figure 2 shows how EDUs are organized into a hierarchical tree.[3]

The RST tree parsing is formalized as $P(t_i \mid c_i;\theta_{\mathcal{A}}) = \prod_{j=1}^{m} P(e_{i_j} \mid c_i;\theta_{\mathcal{A}}) \cdot \prod_{(u,v)} P(r_{u,v} \mid e_{i_u}, e_{i_v};\theta_{\mathcal{A}})$, where $P(e_{i_j} \mid c_i)$ signifies the probability of EDU boundary prediction and $u, v \in V_i$ are discourse units, $P(r_{u,v} \mid e_{i_u}, e_{i_v})$ corresponds to the probability of the rhetorical relation between two EDUs, and $\theta_{\mathcal{A}}$ indicates the parameters of the parser.

Inter-Chunk Rhetorical Graph. For all retrieved chunks $\mathcal{C}$, we construct a directed graph $\mathcal{G} = (\mathcal{C}, \mathcal{F})$. The edge set $\mathcal{F} \subseteq \mathcal{C} \times \mathcal{C} \times (\mathcal{R} \cup \texttt{UNRELATED})$ encodes rhetorical relations or lack thereof. We adopt a listwise inference strategy, where all retrieved chunks $\mathcal{C}$ are provided to parser $\mathcal{A}$ in a single pass, and $\mathcal{A}$ jointly predicts a set of di-

[2] Prompt is detailed in Appendix Figure 10.

[3] Intra-chunk RST trees are constructed offline.

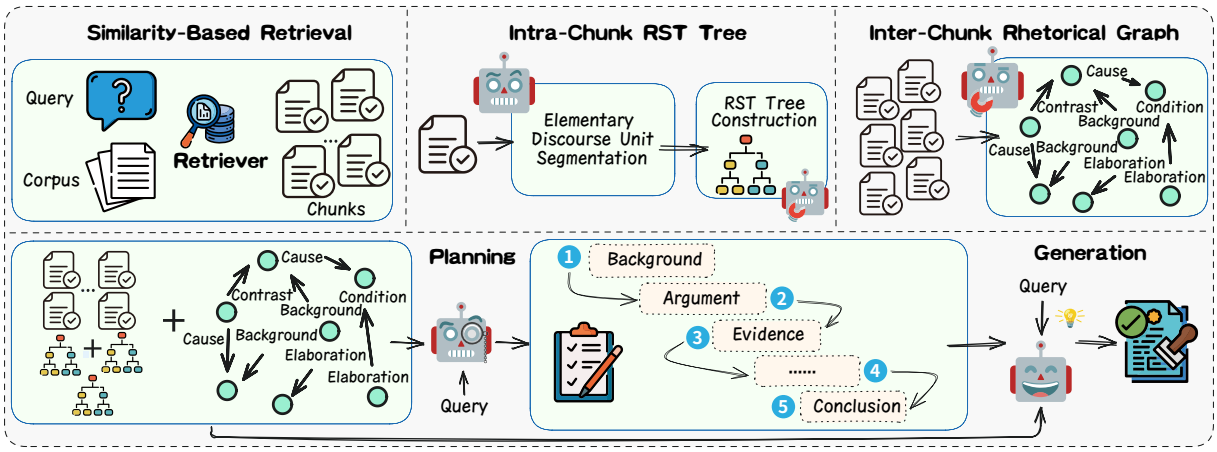

Figure 2: The Disco-RAG pipeline: Starting from passage retrieval (providing context), then intra-chunk RST tree parsing (capturing local discourse), inter-chunk rhetorical graph construction (modeling global discourse), rhetorical planning (blueprint generation), and answer generation (producing the final output).

rected rhetorical relations $\{r_{i,j}\}$ or an UNRELATED label for all chunk pairs.[4]

The rhetorical graph construction is modeled as $P(\mathcal{G} \mid \mathcal{C}; \theta_{\mathcal{A}})$. This joint distribution can be factorized over ordered chunk pairs as $P(\mathcal{G} \mid \mathcal{C}; \theta_{\mathcal{A}}) = \prod_{i=1}^{k} \prod_{j=1, j\neq i}^{k} P(r_{i,j} \mid \mathcal{C}; \theta_{\mathcal{A}})$. As shown in the top-right panel of Figure 2, the resulting graph $\mathcal{G}$ serves as a global discourse scaffold, allowing the generator to reason over cross-chunk connections.

Discourse-Driven Planning. To move beyond the flat concatenation of retrieved evidence, we introduce a planning module that produces a rhetorically informed blueprint to guide the text generation. This is modeled through a mapping from the input query q, retrieved chunks $\mathcal{C}$ together with their RST trees $\mathcal{T}$, and the inter-chunk rhetorical graph $\mathcal{G}$ into a discourse-aware plan $(q, \mathcal{C}, \mathcal{T}, \mathcal{G}) \xrightarrow{\mathcal{A}} \mathcal{B}$.

As depicted in the center-bottom panel of Figure 2, the plan $\mathcal{B}$ is dynamically conditioned on the discourse structures and the query[5]. The plan outlines reasoning steps that involve selecting salient content, organizing argumentative flow, and prioritizing supporting evidence.

Discourse-Guided RAG. The final stage of generation is conditioned on four inputs: (1) the original text chunks $\mathcal{C}$; (2) the intra-chunk RST trees $\mathcal{T}$; (3) the inter-chunk rhetorical graph $\mathcal{G}$; and (4) the discourse-aware plan $\mathcal{B}$. The objective is $y = \arg\max_{y'} P(y' \mid q, \mathcal{C}, \mathcal{T}, \mathcal{G}, \mathcal{B})$, where y' denotes a candidate output and y refers to the final output that maximizes the conditional probability.[6]

4 Experimental Settings

Evaluation Datasets. We evaluate our method on three benchmarks, namely Loong (Wang et al., 2024a), ASQA (Stelmakh et al., 2022), and SciNews (Liu et al., 2024). The Loong dataset focuses on knowledge-intensive reasoning with Spotlight Locating (Spot.), Comparison (Comp.), Clustering (Clus.), and Chain of Reasoning (Chain.). These tasks are evaluated under varying document lengths, where longer inputs increase evidence fragmentation and reasoning difficulty. ASQA involves long-form question answering and requires models to generate responses that are coherent and factually grounded. SciNews targets long-document lay summarization, where the objective is to rewrite scientific articles into accurate and accessible summaries for general audiences (Cachola et al., 2025). Dataset statistics are reported in Appendix Table 6.

Automatic Metrics. To ensure consistency and fair comparison, we follow the official evaluation protocols provided by each dataset's repository (Wang et al., 2024a; Stelmakh et al., 2022; Liu et al., 2024). For the Loong dataset (Wang et al., 2024a; Li et al., 2025b), we report results using Exact Match (EM) and LLM-based scores. For ASQA (Stelmakh et al., 2022; Chang et al., 2025), the evaluation includes EM, ROUGE-L (RL) (Lin, 2004), and DR Score (Stelmakh et al., 2022). On SciNews, we evaluate with RL, BERTScore (Zhang et al., 2020), SARI (Xu et al., 2016), and SummaC (Laban et al., 2022). These metrics assess the informativeness, fluency, and factual consistency of

[4] Appendix Figure 11 provides prompt and format details used in inter-chunk relation prediction.

[5] Appendix Figure 12 provides prompt used in discourse-aware planning.

[6] Appendix Figure 18 contains the generation prompt.

generated answers. Detailed descriptions of these metrics are provided in Appendix B.

Implementation Details. Unless specified otherwise, we use `Llama-3.1-8B`, `Llama-3.3-70B`, or `Qwen2.5-72B` across all modules to instantiate and compare performance at different model scales and families (Grattafiori et al., 2024).[7] For embedding and retrieval modules, we utilize `Qwen3-Embedding-8B` (Zhang et al., 2025c) with a chunk size of 256 tokens without sliding window, and Top-10 retrieval based on cosine semantic similarity. We run each setting once; we use beam search with a beam width of 3, and fix all retrieval settings across compared methods. For Loong and ASQA, retrieval is conducted over the entire corpus, reflecting an open-domain setting. For SciNews, retrieval is restricted to the source document associated with each summary, reflecting a closed-domain setup.

Selected Baselines. We compare `Disco-RAG` against three baseline settings: (1) zero-shot LLMs (`Llama-3.1-8B`, `Llama-3.3-70B`, and `Qwen2.5-72B`) with full input context. (2) standard RAG approach (Lewis et al., 2020)[8], where relevant chunks are prepended to the query prior to inference.[9] and (3) previously published results from state-of-the-art RAG (if applicable) systems on the same benchmark.

5 Results

Main Results. The experimental results are summarized in Table 1, Table 2, and Table 3, which correspond to the Loong, ASQA, and SciNews benchmarks, respectively. Across all benchmarks and evaluation metrics, `Disco-RAG` consistently delivers stable and substantial improvements over the standard RAG baseline.

On the Loong benchmark, `Disco-RAG` demonstrates clear gains across varying document-length settings. With `Llama-3.3-70B` as the backbone, our method achieves an LLM Score of 71.00 in Set 1, outperforming standard RAG by 8.22 points. The performance gap becomes more significant in Set 4, where `Disco-RAG` scores 54.62 compared to 35.61 for standard RAG. Averaged across all four sets, our approach also surpasses the best prior reported training-based method `StructRAG` (62.07 vs. 60.38).

On ASQA, our method again yields consistent advantages. With `Llama-3.1-8B`, EM, RL, and DR scores increase from 37.3/36.9/23.4 to 40.4/42.2/32.6, and with `Llama-3.3-70B`, EM rises to 42.0 and DR to 32.8. Notably, our method outperforms more sophisticated prompting systems, such as `MAIN-RAG` (42.0 RL) and `Tree of Clarifications` (39.7 RL), achieving an RL score of 42.3. On the SciNews summarization task, our approach exhibits strong generalization ability. Using `Llama-3.3-70B`, `Disco-RAG` obtains 21.11 RL score, 65.67 BERTScore, 44.37 SARI, and 69.48 SummaC, surpassing both standard RAG and the previous best system (Liu et al., 2024, 2025b).

Ablation Studies. We perform ablation studies on the Loong benchmark, as summarized in Table 4, to assess the contribution of each component in `Disco-RAG`. We find that the removal of any single module leads to performance degradation. The full model achieves an overall LLM Score of 62.07, which drops to 56.22, 57.10, and 59.75 when the RST tree, rhetorical graph, and planner are removed, respectively. Similarly, the Exact Match metric decreases from 0.24 in the full setting to values ranging from 0.20 to 0.22 across the ablated variants. We also include two generic planning baselines built on standard RAG to isolate the added value of discourse structure modeling beyond planning alone.[10]

Among the three components, the RST tree and rhetorical graph prove to be the most critical. In the long-document setting (Set 4), eliminating the RST tree leads to a decrease in LLM Score from 54.62 to 47.63. Similarly, removing the rhetorical graph reduces the score to 48.16, whereas excluding the planner causes a smaller drop to 50.34. These findings imply that while all three modules contribute complementarily, structural modeling within and across chunks plays a central role in aggregating information and maintaining discourse coherence in long-context generation.

6 Analysis

Impact of Retrieval Granularity and Noise Robustness. To assess the robustness of `Disco-RAG`

[7] `Llama-3.1-8B`, `Llama-3.3-70B`, and `Qwen2.5-72B` are the abbreviated names for `Llama-3.1-8B-Instruct`, `Llama-3.3-70B-Instruct`, and `Qwen2.5-72B-Instruct`.

[8] Prompts for full context generation and standard RAG method are provided in Appendix Figure 13 and Figure 14.

[9] All experiments are training-free and use only task instructions without in-context examples. All hyperparameters follow the same settings as Disco-RAG.

[10] Prompts for these two generic planning baselines are provided in Appendix Figure 15 and Figure 16.

Condition	Model	Spot.		Comp.		Clus.		Chain.		Overall	
		LLM Score↑	EM↑	LLM Score↑	EM↑	LLM Score↑	EM↑	LLM Score↑	EM↑	LLM Score↑	EM↑
Set 1 (10K–50K Tokens)											
Full Context	Llama-3.1-8B	55.43	0.35	56.06	0.36	47.41	0.08	65.66	0.37	56.16	0.30
	Qwen2.5-72B	55.11	0.34	57.21	0.33	47.09	0.10	66.51	0.36	56.59	0.31
	Llama-3.3-70B	58.82	0.44	61.33	0.35	48.15	0.11	**70.31**	0.37	59.54	0.32
Standard RAG	Llama-3.1-8B	62.61	0.32	60.61	0.26	53.61	0.08	58.76	0.32	60.08	0.25
	Qwen2.5-72B	63.20	0.32	61.29	0.35	54.14	0.11	64.67	0.34	61.58	0.33
	Llama-3.3-70B	68.44	0.45	65.32	0.39	55.30	0.12	66.48	0.36	62.78	0.34
SOTA Results	RQ-RAG* (Chan et al., 2024)	72.31	**0.54**	48.16	0.05	47.44	0.07	58.96	0.25	53.51	0.17
	GraphRAG* (Edge et al., 2024)	31.67	0.00	27.60	0.00	40.71	0.14	54.29	**0.43**	40.82	0.18
	StructRAG (Li et al., 2025b)	74.53	0.47	75.58	**0.47**	65.13	**0.23**	67.84	0.34	69.43	0.35
	Disco-RAG (Llama-3.1-8B)	73.35	0.40	73.57	0.37	64.44	0.12	68.00	0.34	69.18	0.32
	Disco-RAG (Qwen2.5-72B)	74.46	0.42	74.39	0.41	64.66	0.15	67.73	0.35	69.59	0.36
	Disco-RAG (Llama-3.3-70B)	**76.60**	0.45	**75.65**	0.45	**65.36**	0.17	68.30	0.38	**71.00**	**0.38**
Set 2 (50K–100K Tokens)											
Full Context	Llama-3.1-8B	51.30	0.27	42.37	0.21	38.32	0.06	44.49	0.11	43.78	0.14
	Qwen2.5-72B	52.37	0.30	44.47	0.25	39.24	0.07	47.69	0.11	46.61	0.13
	Llama-3.3-70B	55.27	0.34	47.93	0.26	40.05	0.08	50.08	0.10	48.24	0.17
Standard RAG	Llama-3.1-8B	57.02	0.25	45.42	0.19	44.21	0.05	50.42	0.15	49.12	0.16
	Qwen2.5-72B	60.13	0.26	50.64	0.20	45.17	0.05	53.28	0.16	50.33	0.17
	Llama-3.3-70B	60.38	0.27	53.37	0.22	45.76	0.07	56.73	0.18	53.77	0.18
SOTA Results	RQ-RAG* (Chan et al., 2024)	57.35	0.35	50.83	0.16	42.85	0.03	47.60	0.10	47.09	0.10
	GraphRAG* (Edge et al., 2024)	24.80	0.00	14.29	0.00	37.86	0.00	46.25	0.12	33.06	0.03
	StructRAG (Li et al., 2025b)	68.00	**0.41**	63.71	**0.36**	61.40	0.17	54.70	0.19	60.95	0.24
	Disco-RAG (Llama-3.1-8B)	66.03	0.36	63.56	0.24	59.53	0.14	53.06	0.16	59.03	0.23
	Disco-RAG (Qwen2.5-72B)	67.17	0.36	64.06	0.30	60.63	0.15	57.22	0.20	61.32	0.25
	Disco-RAG (Llama-3.3-70B)	**69.92**	0.39	**64.34**	**0.36**	**61.67**	**0.18**	**58.23**	**0.22**	**63.61**	**0.28**
Set 3 (100K–200K Tokens)											
Full Context	Llama-3.1-8B	42.25	0.22	37.43	0.12	32.27	0.00	35.62	0.00	36.51	0.08
	Qwen2.5-72B	45.47	0.29	40.13	0.13	35.29	0.01	48.47	0.01	42.01	0.10
	Llama-3.3-70B	47.31	0.31	41.11	0.14	35.64	0.01	49.78	0.01	42.27	0.11
Standard RAG	Llama-3.1-8B	49.22	0.21	40.24	0.03	36.04	0.00	49.05	0.00	43.42	0.06
	Qwen2.5-72B	50.14	0.25	41.83	0.04	40.07	0.03	49.09	0.02	44.38	0.11
	Llama-3.3-70B	50.33	0.33	43.70	0.06	40.13	0.04	50.10	0.05	45.77	0.13
SOTA Results	RQ-RAG* (Chan et al., 2024)	50.50	0.13	44.62	0.00	36.98	0.00	36.79	0.07	40.93	0.05
	GraphRAG* (Edge et al., 2024)	15.83	0.00	27.40	0.00	42.50	0.00	43.33	**0.17**	33.28	0.04
	StructRAG (Li et al., 2025b)	**68.62**	**0.44**	57.74	**0.35**	58.27	**0.10**	49.73	0.13	57.92	0.21
	Disco-RAG (Llama-3.1-8B)	60.76	0.26	55.80	0.11	53.07	0.05	50.31	0.08	56.64	0.15
	Disco-RAG (Qwen2.5-72B)	65.58	0.33	56.89	0.19	57.23	0.06	51.20	0.13	57.14	0.18
	Disco-RAG (Llama-3.3-70B)	66.37	0.38	**57.84**	0.28	**58.85**	0.07	**52.17**	0.15	**58.86**	**0.22**
Set 4 (200K–250K Tokens)											
Full Context	Llama-3.1-8B	31.79	0.12	25.37	0.06	27.87	0.00	26.76	0.00	27.82	0.04
	Qwen2.5-72B	34.22	0.18	28.23	0.06	28.11	0.00	28.48	0.00	30.15	0.04
	Llama-3.3-70B	36.76	0.21	32.22	0.07	30.69	0.00	30.17	0.00	32.21	0.05
Standard RAG	Llama-3.1-8B	40.01	0.11	31.90	0.00	32.33	0.00	29.92	0.00	33.52	0.02
	Qwen2.5-72B	40.14	0.16	32.31	0.01	34.00	0.00	30.02	0.01	33.64	0.04
	Llama-3.3-70B	40.27	0.25	34.49	0.02	36.41	0.01	31.33	0.02	35.61	0.07
SOTA Results	RQ-RAG* (Chan et al., 2024)	29.17	0.08	40.36	0.00	26.92	0.00	34.69	0.00	31.91	0.01
	GraphRAG* (Edge et al., 2024)	17.50	0.00	26.67	0.00	20.91	0.00	33.67	**0.33**	23.47	0.05
	StructRAG (Li et al., 2025b)	56.87	0.19	55.62	**0.25**	56.59	0.00	35.71	0.05	51.42	0.10
	Disco-RAG (Llama-3.1-8B)	56.68	0.19	53.92	0.12	**57.53**	0.02	36.00	0.03	50.87	0.08
	Disco-RAG (Qwen2.5-72B)	57.27	0.22	54.97	0.15	57.40	0.02	**36.17**	0.06	54.47	0.10
	Disco-RAG (Llama-3.3-70B)	**57.74**	**0.24**	**55.80**	0.17	57.36	**0.03**	36.06	0.06	**54.62**	**0.11**

Table 1: Loong benchmark results across four document-length settings. Our method (`Disco-RAG`) is compared against zero-shot LLMs with full context, standard RAG, and prior SOTA. * means that the results are directly taken from Li et al. (2025b). We use **bold red** to indicate the best results and blue underlined text to indicate the second-best results.

under different retrieval conditions, we execute a series of controlled experiments that manipulate the chunk size of passages, the number of Top-k retrieved chunks, and the proportion of noisy passages. All experiments are conducted on the Loong dataset using `Llama-3.3-70B` as the generator model. We maintain identical prompts and decoding configurations across all systems. The evaluation includes two baseline methods, namely the full context setting and the standard RAG framework. Performance is reported using the average LLM Score over four subsets, and the results are visualized in Figure 3.

Panel (a) of Figure 3 shows that standard RAG performs best at a chunk size of 256 tokens (49.33) but degrades with larger chunks due to the loss of structural coherence. In contrast, `Disco-RAG` maintains stable performance across all chunk sizes, with scores ranging from 62.07 to 58.94, showing strong robustness to granularity shifts. Panel (b) of Figure 3 shows that while standard RAG peaks at Top-10 and declines with larger k due to accumulating noise, `Disco-RAG` also performs best at Top-10 but remains robust up to Top-50, showing enhanced capacity to integrate and filter redundant information. Panel (c) of Figure 3 evaluates noise robustness by replacing a fraction of the Top-10 retrieved passages with unrelated content. We randomly replace a proportion of retrieved chunks (*e.g.,* 20%, 40%) with irrelevant ones sampled at random from a pool of non-retrieved chunks. The standard RAG exhibits a steep performance drop from 49.33 to 45.23 as noise increases, whereas `Disco-RAG` retains a score of 56.17, highlighting the structural resilience of our method to retrieval errors.

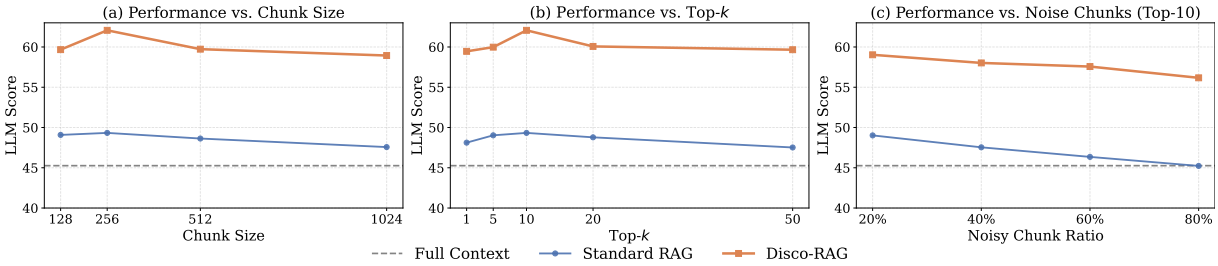

Figure 3: Performance comparison under varying chunk size (a), Top-k value (b), and retrieval noise level (c).

Model	EM↑	RL↑	DR Score↑
Baselines with full context			
Llama-3.1-8B	20.1	30.6	16.3
Qwen2.5-72B	21.3	31.8	17.1
Llama-3.3-70B	22.7	32.9	16.8
Baselines with standard RAG			
Llama-3.1-8B	37.3	36.9	23.4
Qwen2.5-72B	37.7	37.2	23.7
Llama-3.3-70B	38.2	37.2	24.1
SOTA Results			
FLARE (Jiang et al., 2023)	41.3	34.3	31.1
Tree of Clarifications (Kim et al., 2023)	—	39.7	36.6
Open-RAG (Islam et al., 2024)	36.3	38.1	—
ConTReGen (Roy et al., 2024)	41.2	—	30.3
DualRAG (Cheng et al., 2025)	—	31.7	—
RAS (Jiang et al., 2025)	—	39.1	—
MAIN-RAG-Mistral-7B (Chang et al., 2025)	35.7	36.2	—
MAIN-RAG-Llama3-8B (Chang et al., 2025)	39.2	42.0	—
Ours			
Disco-RAG (Llama-3.1-8B)	40.4	42.2	32.6
Disco-RAG (Qwen2.5-72B)	41.8	41.3	**33.2**
Disco-RAG (Llama-3.3-70B)	**42.0**	**42.3**	32.8

Table 2: Performance on the ASQA benchmark. `Disco-RAG` consistently outperforms standard RAG across all metrics. It also surpasses existing SOTA methods on most dimensions.

Model	RL↑	BERTScore↑	SARI↑	SummaC↑
Baselines with full context				
Llama-3.1-8B	15.33	59.27	35.43	48.31
Qwen2.5-72B	17.00	60.41	37.62	55.03
Llama-3.3-70B	17.19	61.03	37.65	54.73
Baselines with standard RAG				
Llama-3.1-8B	17.12	60.35	38.01	55.26
Qwen2.5-72B	18.09	61.28	38.32	60.12
Llama-3.3-70B	18.17	61.37	37.74	60.39
SOTA Results				
RSTformer (Liu et al., 2024)	20.12	62.80	41.56	—
SingleTurnPlan (Liang et al., 2024)	19.68	—	—	—
Plan-Input (Liu et al., 2025b)	—	65.32	—	**72.40**
Ours				
Disco-RAG (Llama-3.1-8B)	19.25	63.47	40.25	63.35
Disco-RAG (Qwen2.5-72B)	20.10	64.83	41.48	66.30
Disco-RAG (Llama-3.3-70B)	**21.11**	**65.67**	**44.37**	69.48

Table 3: Performance on the SciNews dataset. `Disco-RAG` beats both zero-shot and standard RAG, and often surpasses prior SOTA across multiple metrics.

Impact of Structure Quality and Perturbation Analysis. To determine whether the performance gains of `Disco-RAG` arise from the quality of structural modeling rather than the mere presence of structural cues, we conduct a set of controlled perturbation experiments targeting three core components of our framework. These include intra-chunk RST trees, inter-chunk rhetorical graphs, and discourse-aware plans. For each module, we introduce partial degradations by randomly selecting relation labels, edge directions, or planning steps, and either replacing or removing them. This design ensures that the perturbed structures still retain partial coherence, allowing us to assess how sensitive the model is to incomplete or noisy signals. All experiments are conducted with `Llama-3.3-70B` under consistent retrieval and decoding conditions to maintain causal interpretability.

Figure 4 presents the results of the perturbation study. Panel (a) of Figure 4 exhibits that perturbing intra-chunk structures leads to consistent performance decrease. Randomly shuffling a portion of rhetorical relation labels reduces the LLM Score from 62.07 to 55.48. Randomly altering some nucleus–satellite roles lowers the score to 55.15. Removing a randomly selected subtree connection decreases the score to 56.77. Panel (b) of Figure 4 presents the effect of modifying rhetorical graphs. Randomly removing some graph connections between chunks reduces the score to 57.60. Randomly flipping the directions of a subset of edges yields 55.82, while replacing some discourse relation labels within the graph gives 55.50. Panel (c) of Figure 4 analyzes the degradation of rhetorical plans. Omitting the plan altogether reduces performance to 59.75. Shuffling some of the step sequences causes a decline to 57.50, while removing a subset of steps results in 58.14.

Across all three dimensions, structural perturbations lead to performance reduction, yet do not eliminate the benefits conferred by structure-aware modeling. Even when exposed to corrupted or incomplete signals, `Disco-RAG` consistently outperforms both the standard RAG and the full context setting. These results confirm that the observed improvements are not merely due to the inclusion of additional tokens, but instead arise from the model's capacity to leverage structural signals.

Human Evaluation. We conduct a human evaluation on the SciNews dataset. We randomly sample 15 test articles and ask three graduate students

Method	Set 1		Set 2		Set 3		Set 4		Overall	
	LLM Score↑	EM↑	LLM Score↑	EM↑	LLM Score↑	EM↑	LLM Score↑	EM↑	LLM Score↑	EM↑
Disco-RAG (full)	**71.00**	**0.38**	**63.61**	**0.28**	**58.86**	**0.22**	**54.62**	**0.11**	**62.07**	**0.24**
w/o RST tree	65.45	0.34	58.41	0.22	54.90	0.14	47.63	0.07	56.22	0.20
w/o rhetorical graph	67.80	0.33	58.87	0.24	54.04	0.15	48.16	0.10	57.10	0.21
w/o planning	69.11	0.35	60.14	0.25	57.20	0.20	50.34	**0.12**	59.75	0.22
Standard RAG	62.78	0.34	53.77	0.18	45.77	0.13	35.61	0.07	49.33	0.17
w/ retrieve-and-plan	64.05	0.35	54.92	0.18	46.11	0.14	37.22	0.07	50.64	0.14
w/ plan-and-retrieve	64.62	0.35	55.38	0.19	47.82	0.14	38.08	0.08	51.38	0.18

Table 4: Ablation study of the three modules in Disco-RAG with Llama-3.3-70B. *w/o RST tree* removes intra-chunk modeling, *w/o rhetorical graph* removes inter-chunk modeling, and *w/o planning* removes discourse-aware planning. We additionally report two generic planning baselines built on standard RAG. *retrieve-and-plan* generates a free-form plan conditioned on retrieved chunks before generation, and *plan-and-retrieve* first generates a free-form plan from the query and then performs a retrieval step guided by this plan.

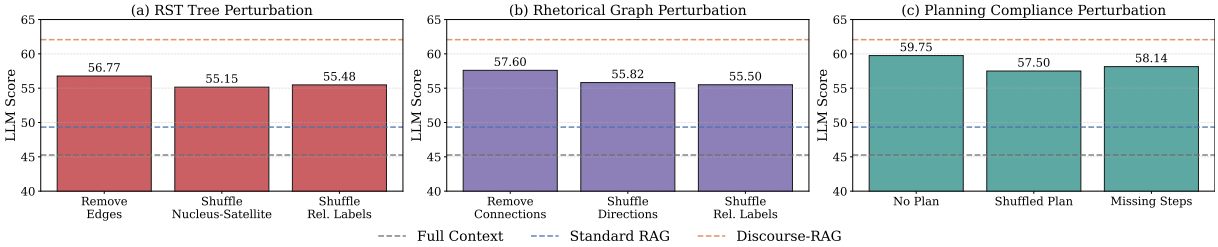

Figure 4: Effect of structural perturbations on performance. Panels (a), (b), and (c) correspond to intra-chunk RST trees, inter-chunk rhetorical graphs, and discourse-aware plans, respectively. Each perturbation involves randomly altering or removing the relevant elements.

with computer science backgrounds to rate four anonymized systems, namely the full context LLM without retrieval, the standard RAG baseline, our Disco-RAG model, and human-written references. Following the protocol of Liu et al. (2024), human raters read each article together with four shuffled summaries and assign scores on a three-point Likert scale along four dimensions, *Relevance*, *Simplicity*, *Conciseness*, and *Faithfulness*, where higher values indicate better quality. We measure inter-rater agreement using Fleiss' κ and obtain average values of 0.72, 0.65, 0.66, and 0.68 on the four dimensions, indicating substantial consistency among annotators. Table 5 reports the average scores across all annotated samples. Detailed instructions for the human raters are provided in Appendix M.

System	Relevance↑	Simplicity↑	Conciseness↑	Faithfulness↑
Full Context	1.65	1.98	1.52	1.45
Standard RAG	1.87	2.12	1.60	1.67
Disco-RAG	2.40	2.43	2.27	2.53
Human Reference	2.89	2.63	2.48	2.88

Table 5: Average human ratings on SciNews. Scores are computed on a three-point Likert scale, and higher values indicate better performance.

Table 5 suggests that Disco-RAG improves perceived answer quality over both full context and standard RAG systems, with considerable gains in *Faithfulness* and *Conciseness*. Human-written references remain the strongest overall according to annotators, which indicates that there is still room for future model development, but the ranking of neural systems in human evaluation is consistent with the trends observed in automatic metrics and supports the benefits of discourse-aware retrieval-augmented generation. Further discussion of the parsing evaluation, efficiency analysis, shallow discourse marker analysis, the impact of model training, significance testing, attention and decoding behavior, qualitative case studies, and LLM usage can be found in Appendix D, Appendix E, Appendix F, Appendix G, Appendix H, Appendix I, Appendix J and Appendix K, respectively.

7 Conclusion

In this study, we tackle the absence of discourse structure modeling in existing RAG approaches by presenting Disco-RAG. Grounded in Rhetorical Structure Theory, our approach constructs both local hierarchies and global discourse representations over retrieved evidence and leverages them to derive a high-level blueprint that guides the reasoning process of the language model. Experimental results demonstrate that Disco-RAG achieves considerable gains across multiple knowledge-intensive QA and summarization tasks, surpassing previous state-of-the-art methods without in-domain fine-tuning. Ablation studies validate the complementary contributions of each structural component. Taken together, these findings highlight structured discourse modeling as a promising direction for advancing retrieval-augmented generation.

8 Ethical Considerations

All datasets used in this work are publicly available, and we follow the original licenses and usage policies. Our pipeline operates entirely on de-identified text without collecting or inferring personal identities or sensitive attributes, and all intermediate artifacts, such as discourse structures and plans, are derived solely from these corpora rather than live user queries or proprietary logs. Human evaluators participate voluntarily and are appropriately compensated, while care is taken to avoid exposing annotators to harmful content beyond what already exists in the datasets. Large language models are used as backbone retrievers/generators and as assistive tools for discourse parsing and language refinement (as noted in Appendix K), but they do not replace the authors in methodological design or result interpretation. We also comply with ACL Policy on Publication Ethics, and we caution against applying our system in high-stakes environments without additional safeguards and human oversight.

9 Limitations

Data. Our experiments apply three publicly available benchmarks, Loong, ASQA, and SciNews. These datasets provide a good basis for evaluating long-context reasoning, but we do not conduct a dedicated analysis of potential biases in their content or label distributions. As a result, the behavior of `Disco-RAG` across genres, languages, or data collection processes that differ substantially from these benchmarks remains an open question, and extending our analysis to different corpora is a direction for future work.

Model. We instantiate our framework with three open-source large language models, including `Llama-3.1-8B`, `Llama-3.3-70B`, and `Qwen2.5-72B`, together with a fixed retriever. The consistent gains observed across these backbones suggest that the proposed discourse mechanism is not tied to a specific model, yet we do not systematically explore alternative architectures, parameter scales, or decoding strategies. In practice, our framework assumes that the backbone models provide basic discourse understanding and long-context processing, and we expect that future improvements in foundation models can be incorporated with minimal changes to the overall design so that `Disco-RAG` remains largely decoupled from specific language models. Further work is also needed to understand how `Disco-RAG` behaves with smaller or more specialized models and under tighter computational constraints.

Parser. Our method depends on an LLM-based discourse parser to produce intra-chunk RST trees and inter-chunk rhetorical graphs. In Appendix D, we evaluate this parser on the standard RST-DT benchmark and show that the zero-shot LLM parser attains span and nuclearity F1 scores that are close to a fine-tuned supervised baseline. Together with the ablation and perturbation studies in Table 4 and Figure 4, these results indicate that better discourse structures lead to stronger downstream performance, and more accurate parsers can be readily plugged into our pipeline without changing the retrieval or generation components. Our goal in this work is not to combine `Disco-RAG` with the strongest available parser, but rather to demonstrate that explicitly modeling discourse information is beneficial for RAG. Since the Loong, ASQA, and SciNews datasets do not provide gold discourse annotations, we cannot directly quantify parsing accuracy on data samples, and we leave more fine-grained comparisons with alternative parsers and discourse formalisms to future work.

Automated Evaluation. We evaluate models using a combination of Exact Match, ROUGE-L, DR Score, BERTScore, SARI, SummaC, and an LLM-based metric for Loong that relies on `GPT-4-turbo-2024-04-09`. This suite covers multiple aspects of quality and has been adopted in prior work, while each metric has known limitations, and the LLM-based judge may inherit biases or topic preferences from its own training data. Reported numbers should therefore be interpreted as indicative rather than exhaustive, and future work could benefit from more fine-grained evaluation protocols and larger-scale human studies.

Efficiency. Compared with standard RAG, `Disco-RAG` introduces additional token consumption and latency because it requires additional LLM calls for discourse parsing and rhetorical planning. The measurements in Appendix E show that this overhead is moderate under our settings. Deploying our method in latency-sensitive or large-scale applications will therefore require engineering optimizations such as caching and reusing discourse structures, batching structural queries, or distilling lighter parsers and planners, and there remains an inherent trade-off between

structural richness and runtime efficiency.

Scope and Generalization. The present work focuses on long-context question answering and summarization. The improvements we observe on Loong, ASQA, and SciNews suggest that discourse modeling is beneficial across multiple tasks, but we do not explore applications like dialog-style question answering, interactive agents, multilingual retrieval, or domains with stronger constraints. In our framework, Rhetorical Structure Theory only serves as a knowledge prior for organizing retrieved evidence rather than a claim that it is the only correct formalization of discourse. Investigating how `Disco-RAG` behaves in these broader settings and how different notions of discourse structure influence retrieval-augmented generation is an important direction for future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rilwan Adewoyin, Ritabrata Dutta, and Yulan He. 2022. RSTGen: Imbuing fine-grained interpretable control into long-FormText generators. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1822–1835, Seattle, United States. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal. Association for Computational Linguistics.

Eric J Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. 2025. Forking paths in neural text generation. In *The Thirteenth International Conference on Learning Representations*.

Isabel Cachola, Daniel Khashabi, and Mark Dredze. 2025. Evaluating the evaluators: Are readability metrics good measures of readability? *arXiv preprint arXiv:2508.19221*.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. RQ-RAG: Learning to refine queries for retrieval augmented generation. In *First Conference on Language Modeling*.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. 2025. MAIN-RAG: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2607–2622, Vienna, Austria. Association for Computational Linguistics.

Rong Cheng, Jinyi Liu, Yan Zheng, Fei Ni, Jiazhen Du, Hangyu Mao, Fuzheng Zhang, Bo Wang, and Jianye Hao. 2025. DualRAG: A dual-process approach to integrate reasoning and retrieval for multi-hop question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31877–31899, Vienna, Austria. Association for Computational Linguistics.

Elena Chistova. 2023. End-to-end argument mining over varying rhetorical structures. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3376–3391, Toronto, Canada. Association for Computational Linguistics.

Elena Chistova. 2024. Bilingual rhetorical structure parsing with large parallel annotations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9689–9706, Bangkok, Thailand. Association for Computational Linguistics.

Song Duong, Florian Le Bronnec, Alexandre Allauzen, Vincent Guigue, Alberto Lumbreras, Laure Soulier, and Patrick Gallinari. 2025. SCOPE: A self-supervised framework for improving faithfulness in conditional text generation. In *The Thirteenth International Conference on Learning Representations*.

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

Masoomali Fatehkia, Ji Kim Lucas, and Sanjay Chawla. 2024. T-rag: lessons from the llm trenches. *arXiv preprint arXiv:2402.07483*.

Akash Gautam, Lukas Lange, and Jannik Strötgen. 2024. Discourse-aware in-context learning for temporal expression normalization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 306–315, Mexico City, Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten,

Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From RAG to memory: Non-parametric continual learning for large language models. In *Forty-second International Conference on Machine Learning*.

Katsuhiko Hayashi, Tsutomu Hirao, and Masaaki Nagata. 2016. Empirical comparison of dependency conversions for RST discourse trees. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 128–136, Los Angeles. Association for Computational Linguistics.

Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2025. GRAG: Graph retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4145–4157, Albuquerque, New Mexico. Association for Computational Linguistics.

Haoyu Huang, Yongfeng Huang, Yang Junjie, Zhenyu Pan, Yongqiang Chen, Kaili Ma, Hongzhi Chen, and James Cheng. 2025. Retrieval-augmented generation with hierarchical knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6044–6060, Suzhou, China. Association for Computational Linguistics.

Shayekh Bin Islam, Md Asib Rahman, K S M Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. 2024. Open-RAG: Enhanced retrieval augmented reasoning with open-source large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14231–14244, Miami, Florida, USA. Association for Computational Linguistics.

Pengcheng Jiang, Lang Cao, Ruike Zhu, Minhao Jiang, Yunyi Zhang, Jimeng Sun, and Jiawei Han. 2025. Ras: Retrieval-and-structuring for knowledge-intensive llm generation. *arXiv preprint arXiv:2502.10996*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 996–1009, Singapore. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Dosung Lee, Wonjun Oh, Boyoung Kim, Minyoung Kim, Joonsuk Park, and Paul Hongsuck Seo. 2025a. ReSCORE: Label-free iterative retriever training for multi-hop question answering with relevance-consistency supervision. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 341–359, Vienna, Austria. Association for Computational Linguistics.

Meng-Chieh Lee, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N. Ioannidis, Huzefa Rangwala, and Christos Faloutsos. 2025b. HybGRAG: Hybrid retrieval-augmented generation on textual and relational knowledge bases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 879–893, Vienna, Austria. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Mufei Li, Siqi Miao, and Pan Li. 2025a. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In *The Thirteenth International Conference on Learning Representations*.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2025b. StructRAG: Boosting knowledge intensive reasoning of LLMs via inference-time hybrid information structurization. In *The Thirteenth International Conference on Learning Representations*.

Lei Liang, Zhongpu Bo, Zhengke Gui, Zhongshu Zhu, Ling Zhong, Peilong Zhao, Mengshu Sun, Zhiqiang Zhang, Jun Zhou, Wenguang Chen, Wen Zhang, and Huajun Chen. 2025. Kag: Boosting llms in professional domains via knowledge augmented generation. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 334–343, New York, NY, USA. Association for Computing Machinery.

Yi Liang, You Wu, Honglei Zhuang, Li Chen, Jiaming Shen, Yiling Jia, Zhen Qin, Sumit Sanghai, Xuanhui Wang, Carl Yang, and 1 others. 2024. Integrating planning into single-turn long-form text generation. *arXiv preprint arXiv:2410.06203*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Teng Lin, Yizhang Zhu, Yuyu Luo, and Nan Tang. 2025. Srag: Structured retrieval-augmented generation for multi-entity question answering over wikipedia graph. *arXiv preprint arXiv:2503.01346*.

Dongqi Liu and Vera Demberg. 2024. RST-LoRA: A discourse-aware low-rank adaptation for long document abstractive summarization. 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 2200–2220, Mexico City, Mexico. Association for Computational Linguistics.

Dongqi Liu, Yifan Wang, and Vera Demberg. 2023. Incorporating distributions of discourse structure for long document abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5574–5590, Toronto, Canada. Association for Computational Linguistics.

Dongqi Liu, Yifan Wang, Jia Loy, and Vera Demberg. 2024. SciNews: From scholarly complexities to public narratives – a dataset for scientific news report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14429–14444, Torino, Italia. ELRA and ICCL.

Dongqi Liu, Chenxi Whitehouse, Xi Yu, Louis Mahon, Rohit Saxena, Zheng Zhao, Yifu Qiu, Mirella Lapata, and Vera Demberg. 2025a. What is that talk about? a video-to-text summarization dataset for scientific presentations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6187–6210, Vienna, Austria. Association for Computational Linguistics.

Dongqi Liu, Xi Yu, Vera Demberg, and Mirella Lapata. 2025b. Explanatory summarization with discourse-driven planning. *Transactions of the Association for Computational Linguistics*, 13:1146–1170.

Guanran Luo, Zhongquan Jian, Wentao Qiu, Meihong Wang, and Qingqiang Wu. 2025. DTCRS: Dynamic tree construction for recursive summarization. In *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.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.

Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, and Manabu Okumura. 2024. Can we obtain significant success in RST discourse parsing by using large language models? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2803–2815, St. Julian's, Malta. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical report, University of Southern California, Information Sciences Institute Los Angeles.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Daniel Marcu. 1997. From discourse structures to text summaries. In *Intelligent Scalable Text Summarization*.

Daniel Marcu. 1999. A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 365–372.

Costas Mavromatis and George Karypis. 2025. GNN-RAG: Graph neural retrieval for efficient large language model reasoning on knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16682–16699, Vienna, Austria. Association for Computational Linguistics.

Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, and 1 others. 2025. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond n-grams: Rethinking evaluation metrics and strategies for multilingual abstractive summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19019–19035, Vienna, Austria. Association for Computational Linguistics.

Hellina Hailu Nigatu, Min Li, Maartje Ter Hoeve, Saloni Potdar, and Sarah Chasins. 2025. mRAKL: Multilingual retrieval-augmented knowledge graph construction for low-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13072–13089, Vienna, Austria. Association for Computational Linguistics.

Renyi Qu, Ruixuan Tu, and Forrest Sheng Bao. 2025. Is semantic chunking worth the computational cost? In *Findings of the Association for Computational*

Linguistics: NAACL 2025, pages 2155–2177, Albuquerque, New Mexico. Association for Computational Linguistics.

Haoran Que and Wenge Rong. 2025. PIC: Unlocking long-form text generation capabilities of large language models via position ID compression. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6982–6995, Vienna, Austria. Association for Computational Linguistics.

Kashob Kumar Roy, Pritom Saha Akash, Kevin Chen-Chuan Chang, and Lucian Popa. 2024. ConTReGen: Context-driven tree-structured retrieval for open-domain long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13773–13784, Miami, Florida, USA. Association for Computational Linguistics.

Diego Sanmartin. 2024. Kg-rag: Bridging the gap between knowledge and creativity. *arXiv preprint arXiv:2405.12035*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. 2025. Parametric retrieval augmented generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1240–1250.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jiaan Wang, Fandong Meng, Zengkui Sun, Yunlong Liang, Yuxuan Cao, Jiarong Xu, Haoxiang Shi, and Jie Zhou. 2025a. An empirical study of many-to-many summarization with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11344, Vienna, Austria. Association for Computational Linguistics.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024a. Leave no document behind: Benchmarking long-context LLMs with extended multi-doc QA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646, Miami, Florida, USA. Association for Computational Linguistics.

Shu Wang, Yixiang Fang, Yingli Zhou, Xilin Liu, and Yuchi Ma. 2025b. Archrag: Attributed community-based hierarchical retrieval-augmented generation. *arXiv preprint arXiv:2502.09891*.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2024b. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736, Miami, Florida, USA. Association for Computational Linguistics.

Zhitong Wang, Cheng Gao, Chaojun Xiao, Yufei Huang, Shuzheng Si, Kangyang Luo, Yuzhuo Bai, Wenhao Li, Tangjian Duan, Chuancheng Lv, Guoshan Lu, Gang Chen, Fanchao Qi, and Maosong Sun. 2025c. Document segmentation matters for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8063–8075, Vienna, Austria. Association for Computational Linguistics.

Jian Wu, Linyi Yang, Dongyuan Li, Yuliang Ji, Manabu Okumura, and Yue Zhang. 2025a. MMQA: Evaluating LLMs with multi-table multi-hop complex questions. In *The Thirteenth International Conference on Learning Representations*.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025b. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467, Vienna, Austria. Association for Computational Linguistics.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2025. MMed-RAG: Versatile multimodal RAG system for medical vision language models. In *The Thirteenth International Conference on Learning Representations*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. 2025. Inference scaling for long-context retrieval augmented generation. In *The Thirteenth International Conference on Learning Representations*.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. eRST: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, 51(1):23–72.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025a. Personalized text generation with contrastive activation steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7128–7141, Vienna, Austria. Association for Computational Linguistics.

Taolin Zhang, Dongyang Li, Qizhou Chen, Chengyu Wang, and Xiaofeng He. 2025b. BELLE: A bi-level multi-agent reasoning framework for multi-hop question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4184–4202, Vienna, Austria. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiaoming Zhang, Ming Wang, Xiaocui Yang, Daling Wang, Shi Feng, and Yifei Zhang. 2024. Hierarchical retrieval-augmented generation model with rethink for multi-hop question answering. *arXiv preprint arXiv:2408.11875*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025c. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Jihao Zhao, Zhiyuan Ji, Zhaoxin Fan, Hanyu Wang, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025. MoC: Mixtures of text chunking learners for retrieval-augmented generation system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5172–5189, Vienna, Austria. Association for Computational Linguistics.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8912–8924, Albuquerque, New Mexico. Association for Computational Linguistics.

A Details of Datasets

Table 6 summarizes the key statistics of the Loong, ASQA, and SciNews datasets used in our experiments. The Loong dataset is a cross-domain and multi-task benchmark that covers long-text understanding, reasoning, and generation. It is specifically designed to evaluate models' ability to handle long-context inputs and perform comprehensive reasoning. The ASQA (Ambiguous Question Answering) dataset focuses on questions with multiple valid interpretations, providing explanatory responses that evaluate a model's capacity to resolve semantic ambiguity and produce interpretable answers. The SciNews dataset centers on the scientific news domain, spanning a wide range of scientific topics. It contains news articles paired with academic papers and is intended to test models' capacity for long-context news understanding and generation.

B Details of Evaluation Metrics

For the Loong dataset. We report two evaluation metrics. The first is Exact Match (EM), which is a strict measure of the percentage of model predictions that exactly match the ground truth answers. It is a binary measure that assigns a score of one for a perfect match and zero otherwise. The second metric is the LLM Score (Wang et al., 2024a), ranging from 0 to 100. Following the protocol introduced by the dataset authors, we employ `GPT-4-turbo-2024-04-09` as an automated evaluator to rate the overall quality of generated responses. Unlike EM, which captures only factual correctness, the LLM Score provides a holistic evaluation by jointly considering comprehensiveness, clarity, and adherence to instructions, thereby offering a more integrated assessment across multiple dimensions of quality.

For the ASQA dataset. We adopt the standard evaluation suite. The first is Exact Match (EM), defined as before. The second is ROUGE-L (Lin, 2004), an evaluation metric based on the Longest Common Subsequence (LCS). It measures the n-gram overlap between prediction and reference by identifying the longest sequence of words that occurs in both while preserving word order, thereby evaluating the coverage of key information. Given a predicted text $\hat{y}_i$ and a reference text y_i, let $LCS(\hat{y}_i, y_i)$ denote the length of their longest common subsequence. The ROUGE-L recall, precision,

Dataset	Loong				ASQA	SciNews
Split	Set1(10K-50K)	Set2(50K-100K)	Set3(100K-200K)	Set4(200K-250K)	Test	Test
Language	EN, ZH	EN, ZH	EN, ZH	EN, ZH	EN	EN
Test Instances	323	564	481	232	1015	4188

Table 6: Summary statistics of the Loong, ASQA, and SciNews datasets used in our experiments.

and F1 are defined as:

$$R_L = \frac{LCS(\hat{y}_i, y_i)}{|y_i|} \tag{1}$$

$$P_L = \frac{LCS(\hat{y}_i, y_i)}{|\hat{y}_i|} \tag{2}$$

$$F_L = \frac{(1+\beta^2) \cdot R_L \cdot P_L}{R_L + \beta^2 \cdot P_L} \tag{3}$$

where $|y_i|$ and $|\hat{y}_i|$ are the lengths of the reference and predicted texts, respectively, and β is set to one by default to balance recall and precision. In our experiments, we report ROUGE-L F1 (RL).

The third metric is the Disambiguation Recall (DR) Score (Stelmakh et al., 2022), which is specifically designed for ASQA to evaluate whether a prediction covers all possible disambiguated answers present in the reference set. While ROUGE-L cannot distinguish between two fluent but semantically divergent answers, the DR score explicitly evaluates coverage across multiple reference answers. A higher DR score indicates that the generated response captures a larger fraction of the possible interpretations of an ambiguous question. Given multiple reference answers $\mathcal{Y}_i = \{y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(k_i)}\}$ for a query and a generated answer $\hat{y}_i$, the instance-level DR score is defined as:

$$\text{DR}_i = \frac{1}{|\mathcal{Y}_i|} \sum_{j=1}^{|\mathcal{Y}_i|} \mathbf{1}\left[\hat{y}_i \text{ contains the information in } y_i^{(j)}\right] \tag{4}$$

where $\mathbf{1}[\cdot]$ is an indicator function equal to one if the predicted answer includes the content of a reference answer $y_i^{(j)}$, and zero otherwise. The overall DR score across N queries is defined as:

$$\text{DR} = \frac{1}{N} \sum_{i=1}^{N} \text{DR}_i. \tag{5}$$

For the SciNews dataset. We focus on summarization quality using four metrics. The first is ROUGE-L, as defined above. The second is BERTScore (Zhang et al., 2020), which computes semantic similarity between prediction and reference using contextual embeddings from a pre-trained BERT model. The third is SARI (Xu et al., 2016), which assesses the quality of simplification by comparing system outputs against both the source text and the reference texts. SARI explicitly measures the precision and recall of words that are added, deleted, and kept. For a source sentence s_i, a prediction $\hat{y}_i$, and a set of reference simplifications $\mathcal{Y}_i = \{y_i^{(1)}, \ldots, y_i^{(k_i)}\}$, SARI is defined as:

$$\text{SARI} = \frac{1}{3}\Big(\text{Add}_{F_1} + \text{Keep}_{F_1} + \text{Del}_{F_1}\Big) \tag{6}$$

where Add_{F_1}, Keep_{F_1}, and Del_{F_1} denote the F1 scores for added, kept, and deleted n-grams relative to both the source and the reference sets. The fourth metric is SummaC (Laban et al., 2022), a model-based measure of factual consistency. SummaC can be used to determine whether a generated summary is entailed by its source document and detects unsupported or hallucinated content, which is essential for ensuring the reliability of generated text.

C Details of Baselines

Here we describe the baselines used for comparison:

- **Standard RAG** (Lewis et al., 2020) We implement the standard retrieval-augmented generation framework, where a retriever (`Qwen3-Embedding-8B`) retrieves relevant documents and a generator (`Llama-3.1-8B`, `Llama-3.3-70B` or `Qwen2.5-72B`) produces the final answer conditioned on the retrieved context.
- **GraphRAG** (Edge et al., 2024) augments retrieval with a graph-based knowledge representation by constructing a semantic knowledge graph from retrieved passages. It leverages community detection to capture global structures and integrates graph contexts into generation, enabling more accurate and coherent reasoning across documents.
- **RQ-RAG** (Chan et al., 2024) refines queries through explicit rewriting, decomposition, and

disambiguation before retrieval. It trains LLMs end-to-end on a curated dataset with search-augmented supervision, enabling dynamic query refinement and improving both single-hop and multi-hop QA by learning to search only when needed.

- **FLARE** (Jiang et al., 2023) actively decides when and what to retrieve during generation by predicting upcoming sentences and using them as queries to fetch additional documents whenever low-confidence tokens appear.
- **Tree of Clarifications** (Kim et al., 2023) addresses ambiguous questions by recursively constructing a tree of disambiguated questions with retrieval-augmented few-shot prompting, pruning unhelpful branches through self-verification, and generating a long-form answer that covers all valid interpretations.
- **Open-RAG** (Islam et al., 2024) enhances retrieval-augmented reasoning with open-source LLMs by transforming a dense model into a parameter-efficient sparse mixture-of-experts, combining contrastive learning against distractors with hybrid adaptive retrieval.
- **ConTReGen** (Roy et al., 2024) employs a context-driven, tree-structured retrieval framework for open-domain long-form text generation. It performs top-down planning to recursively decompose a query into sub-questions for in-depth retrieval, followed by bottom-up synthesis to integrate information from leaf nodes to the root.
- **DualRAG** (Cheng et al., 2025) introduces a dual-process framework for multi-hop QA, consisting of Reasoning-augmented Querying (RaQ), which identifies knowledge gaps and formulates targeted queries, and progressive Knowledge Aggregation (pKA), which filters and structures retrieved information into a coherent knowledge outline.
- **RAS** (Jiang et al., 2025) interleaves iterative retrieval planning with dynamic construction of query-specific knowledge graphs. It converts retrieved text into factual triples, incrementally builds a structured graph, and conditions generation on the evolving graph.
- **MAIN-RAG** (Chang et al., 2025) is a training-free framework that employs three LLM agents to collaboratively filter and rank retrieved documents. It introduces an adaptive judge bar that dynamically adjusts relevance thresholds based on score distributions, effectively reducing noisy retrievals while preserving relevant information.
- **StructRAG** (Li et al., 2025b) introduces hybrid information structurization for knowledge-intensive reasoning. It employs a hybrid structure router to select the optimal structure type (e.g., table, graph, catalogue), a scattered knowledge structurizer to transform raw documents into structured knowledge, and a structured knowledge utilizer to decompose complex questions and infer accurate answers based on the structured representation.

D Parsing Evaluation

To assess the parser quality on our framework, we evaluate the LLM-based parser used in `Disco-RAG` on the RST-DT benchmark following the evaluation protocol of Maekawa et al. (2024). We compare a fine-tuned RST parser from Maekawa et al. (2024) with our zero-shot parser instantiated with `Llama-3.3-70B`. Both models are evaluated on span F1, nuclearity F1, and relation F1 using the official data splits and scoring scripts of the benchmark, and the results are summarized in Table 7. Our zero-shot parser attains competitive scores that are close in nuclearity and relation prediction and somewhat lower in the span prediction, which still reflects reasonable sensitivity to rhetorical semantics without any task-specific tuning.

Model	Setting	Span F1	Nuclearity F1	Relation F1
Maekawa et al. (2024)	Supervised	79.8	70.4	60.0
Our Parser	Unsupervised	70.4	63.1	58.6

Table 7: Evaluation of the RST parser on the RST-DT benchmark following the protocol of Maekawa et al. (2024).

Note that parser development is not the primary focus of `Disco-RAG`, and these results indicate that the zero-shot LLM parser provides a reasonable structural signal for downstream reasoning. Moreover, the fine-tuned parser only generates output with a specific format, and cannot complete the rhetorical graph prediction between chunks, while the zero-shot parser provides such flexibility.

We further conduct a case study to examine whether the parser outputs are acceptable for the downstream task when gold annotations are unavailable for our benchmarks. For each of Loong, ASQA, and SciNews, we randomly select 10 instances from the test set and run our pipeline to obtain intra-chunk RST trees, inter-chunk rhetor-

ical graphs, and discourse-aware plans for the retrieved evidence. We then ask three human annotators to judge two questions with binary labels. The first question evaluates *whether the predicted discourse structures are broadly acceptable*, meaning that they capture major relations within and between chunks even if they are not perfectly accurate in every detail. The second question evaluates *whether the discourse-aware plan is acceptable*, meaning that it organizes the answer in a reasonable order and reflects the main evidence required by the query.

Across the sampled instances, the average acceptability rates are 0.72 for intra-chunk discourse trees, 0.80 for inter-chunk rhetorical graphs, and 0.93 for discourse-aware plans. The inter-annotator agreement measured by Fleiss' κ is 0.709 for intra-chunk discourse trees, 0.733 for inter-chunk rhetorical graphs, and 0.862 for discourse-aware plans, indicating high consistency among annotators. These results suggest that the parsing outputs and the plans provide usable discourse signals for our framework, and we expect that improved parsing performance would further enhance the reliability of discourse structures and thereby support additional gains in answer quality.

E End-to-End Efficiency Analysis

We provide an end-to-end efficiency comparison between standard RAG and `Disco-RAG` to quantify the additional inference cost introduced by structural modeling. Experiments are conducted on the Loong benchmark using the same configuration. To be specific, we adopt the same retrieval corpus and Top-k retrieval strategy (with chunk size fixed at 256 tokens), the same decoding hyperparameters, and we vary only whether discourse-aware components are enabled. Standard RAG performs a single generation call conditioned on the retrieved context. `Disco-RAG` assumes that RST trees over the corpus have been pre-parsed offline, and at inference time adds one listwise inter-chunk rhetorical graph prediction and one discourse-aware planning call before the final answer generation. All measurements are obtained on a cluster equipped with 32 $\times$ NVIDIA A100 80GB GPUs.

To characterize how inference cost scales with the number of retrieved chunks, we fix the retriever and model configuration on Loong and vary only Top-k $\in \{10, 20, 30, 50\}$. For each setting, we record (1) the average token cost, defined as the total number of prompt input and output tokens across all LLM calls, and (2) the end-to-end latency per query, including retrieval, structure prediction, planning, and final generation. As shown in Table 8, across all Top-k values, the token cost of `Disco-RAG` is roughly 2.2$\times$ that of standard RAG, and the end-to-end latency increases by about two seconds per query on average.

Top-k	Standard RAG (Token Cost)	Disco-RAG (Token Cost)	Standard RAG (Latency)	Disco-RAG (Latency)
10	3.6k	7.9k	7.6s	9.8s
20	6.4k	14.3k	13.8s	15.9s
30	8.2k	18.2k	21.5s	23.8s
50	13.5k	29.6k	32.8s	35.1s

Table 8: End-to-end token cost (input + output) and latency under different Top-k settings on the Loong benchmark. Results are averaged over the same set of queries with identical retriever, generator, and decoding configurations for both Standard RAG and Disco-RAG.

Combining these results with the accuracy improvements reported in the main paper on Loong and the other benchmarks, we observe that `Disco-RAG` incurs a moderate and bounded increase in inference cost in exchange for substantial gains in performance over standard RAG, and it often outperforms existing structure-aware (training-based) RAG methods. We view this trade-off between cost and performance as acceptable in knowledge-intensive applications, where RST tree parsing can be fully amortized offline and reused across queries, while the additional listwise discourse inference and planning incur a stable overhead that yields stronger discourse coherence and factual robustness in the generated answers.

F Comparison with Shallow Discourse Markers

We conduct a study on the Loong to assess whether shallow discourse cues alone can provide comparable benefits to full RST-based modeling. To this end, we design a marker-based variant that constructs inter-chunk links using explicit discourse markers without applying EDU segmentation. We adopt a listwise inference strategy and provide all retrieved chunks to `Llama-3.3-70B` in a single pass, which jointly predicts discourse marker for each ordered chunk pair based on connective cues such as *however*, *but*, *although*, *in contrast*, *therefore*, *because*, *as a result*, and *meanwhile*.[11]

[11]The prompt used for shallow discourse marker inference is provided in Appendix Figure 17.

Method	LLM Score↑	Exact Match↑
Standard RAG	49.33	0.17
w/ Discourse Markers	50.41	0.20
Disco-RAG	62.07	0.24

Table 9: Comparison of standard RAG, a shallow discourse marker variant, and Disco-RAG on the Loong benchmark with Llama-3.3-70B.

Table 9 compares three configurations, namely standard RAG, a shallow variant that augments standard RAG with discourse markers, and the full Disco-RAG model. The marker-based system improves LLM Score from 49.33 to 50.41 and Exact Match from 0.17 to 0.20. However, these gains remain modest compared with the full discourse-aware setting, where Disco-RAG reaches 62.07 LLM Score and 0.24 Exact Match under the same conditions.

G Effect of Supervised Fine-Tuning

We examine how supervised fine-tuning interacts with discourse-aware modeling on the SciNews summarization benchmark. Starting from Llama-3.3-70B, we fine-tune the generator on the SciNews training split with a standard sequence-to-sequence summarization objective and test using RAG setting under three conditions. In the end-to-end baseline, the model is trained using only the raw document summary pairs without any discourse inputs. In the second setting, the model is trained in the same way, but at test time, we augment the inputs with the intra-chunk RST trees, inter-chunk rhetorical graphs, and discourse-aware plans produced by Disco-RAG. In the third setting, both training and inference use the discourse-enriched inputs so that the model can adapt its parameters to the structural signals. For comparison, we also include the original training-free Disco-RAG system that conditions generation on discourse structures via prompting without parameter updates.

Method	RL↑	SummaC↑
End-to-end SFT (no discourse)	20.3	66.8
Disco-RAG (training-free)	21.1	69.5
SFT with test time discourse	22.8	72.3
SFT with train and test discourse	23.3	74.0

Table 10: Impact of supervised fine-tuning (SFT) and discourse conditioning.

All systems share the same retrieval pipeline and decoding configuration as described in the main paper, and we report RL and SummaC on the SciNews test set. Table 10 shows that naive end-to-end fine-tuning improves over the zero-shot standard RAG baselines but remains behind the training-free Disco-RAG. When discourse structures are provided at test time, the fine-tuned model surpasses Disco-RAG, indicating that structural guidance and parameter adaptation bring complementary benefits. When discourse structures are incorporated during both training and inference, we observe further gains in both RL and SummaC. These results confirm that our discourse-aware framework is orthogonal to model training and that injecting discourse information can consistently enhance performance on top of supervised fine-tuning.

H Significance Testing

To assess whether the improvements of Disco-RAG over standard RAG are statistically reliable under the same backbone model and decoding configuration, we conduct paired t-tests on metric scores for every benchmark, every backbone, and every automatic metric. For human evaluation, we apply the same paired t-test on the instance-level average ratings across the three annotators for each criterion. Across all evaluation settings reported in the paper, Disco-RAG is significantly better than standard RAG with $p < 0.05$.

I Attention and Decoding Analysis

To understand how discourse-aware modeling influences the decoding process, we analyze inter-layer attention behavior and factual consistency on the SciNews summarization benchmark. Using Llama-3.3-70B as the generator, we compare configurations that mirror the ablation settings in the main paper: the full Disco-RAG model, variants that remove either intra-chunk RST trees, inter-chunk rhetorical graphs, or discourse-aware plans while keeping other components unchanged, and a standard RAG baseline that conditions generation only on retrieved chunks. For each configuration, we compute inter-layer attention entropy by averaging the token-level cross-layer attention distributions over all decoder layers and attention heads, and we report SummaC scores as a measure of factual consistency with respect to the source articles.

Table 11 summarizes the results. As structural

Model Configuration	Inter-layer Attention Entropy$_{\downarrow}$	SummaC$_{\uparrow}$
Disco-RAG	3.72	69.5
w/o Discourse Plan	4.07	68.6
w/o Intra-chunk RST	4.45	65.8
w/o Inter-chunk Graph	4.53	66.9
Standard RAG	6.21	60.4

Table 11: Inter-layer attention entropy and SummaC on the SciNews benchmark with `Llama-3.3-70B` under different structural configurations. Lower entropy and higher SummaC indicate more focused attention and better factual consistency, respectively.

guidance is removed, inter-layer attention entropy increases, indicating that the model attends less selectively to salient elements such as nucleus spans and key graph nodes. At the same time, SummaC scores decrease, reflecting a loss of factual alignment between generated summaries and their sources. The full `Disco-RAG` model achieves the lowest attention entropy and the highest SummaC score among the discourse-aware variants, while standard RAG exhibits the highest entropy and the lowest factual consistency. Taken together, these findings suggest that discourse structures act as soft constraints on content ordering and focus during decoding, promoting more stable and faithful generation.

J Case Studies

We present qualitative case studies to illustrate the effectiveness of `Disco-RAG` compared to standard RAG. The three cases are shown in Figure 5, Figure 6, and Figure 7.

K Use of Large Language Models

In preparing this paper, we use `GPT-5.2` as a writing assistant for language polishing, grammar correction, and stylistic refinement. The model is not involved in the research ideation, methodology design, experiments, or result interpretation. The authors take full responsibility for the content of the manuscript, including any parts generated with the assistance of `GPT-5.2`. In accordance with conference policy, we confirm that the LLM is not an author of this work and does not bear responsibility for its scientific claims.

L Prompts for Disco-RAG

Appendix Figure 10, Figure 11, Figure 12 and Figure 18 present the prompts used in `Disco-RAG`.

M Human Evaluation Guidelines

Case Study 1

Query. *When did The Lion King debut on Broadway?*
Retrieved Chunks.

```
[Chunk 1] The Lion King is a stage musical ... It debuted on July 8, 1997,
in Minneapolis, Minnesota at the Orpheum Theatre ...

[Chunk 2] ... The musical premiered on Broadway at the New Amsterdam Theatre
in previews on October 15, 1997 ...

[Chunk 3] ... The official Broadway opening was on November 13, 1997,
at the New Amsterdam Theatre ...

[Chunk 4] ... On June 13, 2006, the Broadway production moved to the Minskoff Theatre,
where it is still running ...
```

Standard RAG Output.

> The Lion King debuted on Broadway in 1997. It was first performed in Minneapolis before moving to Broadway, where it became one of the most popular shows.

Disco-RAG Output.

> The Lion King opened on Broadway more than once. It premiered in previews at the New Amsterdam Theatre on October 15, 1997, and officially opened on November 13, 1997. Later, on June 13, 2006, the production moved to the Minskoff Theatre, where it continues to run.

Reference Answer.

> The Lion King opened on Broadway more than once. It premiered on Broadway at the New Amsterdam Theatre in previews on October 15, 1997, then officially opened on Broadway on November 13, 1997. On June 13, 2006, the Broadway production moved to the Minskoff Theatre, another Broadway theater, where it is still running after more than 9,000 performances.

Remark. Standard RAG produces a vague answer that misses key details about the Broadway timeline, providing only the year 1997. In contrast, `Disco-RAG` integrates evidence across all relevant chunks, capturing both the preview and official opening dates as well as the later theater relocation.

Figure 5: Case study comparing standard RAG and `Disco-RAG` on the query "When did The Lion King debut on Broadway?". Our method captures both the preview and official opening as well as the later relocation, while standard RAG gives only a vague year-based answer.

Case Study 2

Query. *Where did Italy win the World Cup in 2006?*

Retrieved Chunks.

```
[Chunk 1] The 2006 FIFA World Cup final was played on July 9, 2006,
between Italy and France ... after a 1-1 draw, Italy won 5-3 on penalties ...

[Chunk 2] ... The final was held at the Olympiastadion in Berlin, Germany,
with an attendance of 69,000 spectators ...

[Chunk 3] Italy claimed their fourth world title, having previously
won in 1934, 1938, and 1982 ...

[Chunk 4] ... France had won their only previous World Cup in 1998,
while Italy had lost finals in 1970 and 1994 ...

[Chunk 5] The match was part of the France-Italy football rivalry,
and marked Italy's first world title in 24 years ...
```

Standard RAG Output.

> Italy won the 2006 World Cup in Munich, Germany, after a penalty shootout against France.

Disco-RAG Output.

> Italy defeated France in the 2006 FIFA World Cup Final, played on July 9, 2006, at the Olympiastadion in Berlin, Germany. After a 1–1 draw, Italy prevailed 5–3 in the penalty shootout.

Reference Answer.

> The Italy national football team defeated France and was declared world champions in the 2006 World Cup Final, the last match of the 2006 FIFA World Cup, at the Olympiastadion stadium in Berlin, Germany, on July 9, 2006. Italy had won three previous finals in 1934, 1938, 1982 and lost in 1970 and 1994, while France had won their only previous final in 1998.

Remark. Standard RAG incorrectly states that the final took place in Munich, likely due to spurious retrieval or shallow aggregation of context. In contrast, `Disco-RAG` integrates evidence across multiple chunks, correctly identifying the Olympiastadion in Berlin as the venue and providing richer historical context.

Figure 6: Case study comparing standard RAG and our proposed `Disco-RAG` on the query "Where did Italy win the World Cup in 2006?". Our method correctly identifies the Olympiastadion in Berlin, while standard RAG produces a factual error.

Case Study 3

Query. *Is Pluto a planet?*
Retrieved Chunks.
`[Chunk 1] Pluto was discovered in 1930 and was long regarded as the ninth planet in the Solar System in many textbooks and popular references.`

`[Chunk 2] In 2006, the International Astronomical Union adopted a formal definition of "planet" and reclassified Pluto as a dwarf planet.`

`[Chunk 3] The 2006 definition requires a planet to orbit the Sun, be nearly round, and clear its orbital neighborhood, which Pluto does not satisfy.`

Standard RAG Output.

Yes. Pluto is the ninth planet in the Solar System.

Disco-RAG Output.

Pluto was historically treated as the ninth planet, but it is currently classified as a dwarf planet. In 2006, the International Astronomical Union adopted a definition of a planet that requires clearing the orbital neighborhood, and Pluto does not meet this criterion.

Reference Answer.

Pluto is not classified as a planet under the current astronomical definition. It was long considered the ninth planet after its discovery in 1930, but the International Astronomical Union reclassified it as a dwarf planet in 2006 because it does not clear its orbital neighborhood.

Remark. Standard RAG commits to an outdated claim from Chunk 1 and produces a categorical answer that conflicts with the current definition. In contrast, `Disco-RAG` uses the `Contrast` relation to avoid merging incompatible statements and uses the `Background` relation to ground the final answer in the relevant criterion, which reduces the risk of hallucinating a definitive but incorrect conclusion under conflicting evidence.

Figure 7: Case study showing how discourse relations affect generation under conflicting evidence. The `Contrast` relation prevents incompatible claims from being merged, and the `Background` relation provides the criterion needed for a faithful answer.

Relation Definitions for Intra-chunk RST Tree Construction

Relation Definitions:

- ELABORATION: Satellite provides additional detail or information about the nucleus.
- EXPLANATION: Satellite explains or clarifies the nucleus content.
- EVIDENCE: Satellite provides evidence or proof for the nucleus claim.
- EXAMPLE: Satellite gives a specific example of the nucleus concept.
- CONTRAST: Satellite presents opposing or contrasting information.
- COMPARISON: Satellite compares two or more entities or concepts.
- CONCESSION: Satellite acknowledges opposing viewpoint while maintaining main claim.
- ANTITHESIS: Satellite presents directly opposite or contradictory information.
- CAUSE: Satellite describes the cause of an event or situation.
- RESULT: Satellite describes the result or consequence of an action.
- CONSEQUENCE: Satellite shows the outcome following from the nucleus.
- PURPOSE: Satellite explains the intended goal or purpose.
- CONDITION: Satellite specifies conditions under which something holds.
- TEMPORAL: Satellite indicates temporal relationship between events.
- SEQUENCE: Satellite shows sequential order of events or actions.
- BACKGROUND: Satellite provides background context or setting.
- CIRCUMSTANCE: Satellite describes circumstances surrounding an event.
- SUMMARY: Satellite summarizes or generalizes the nucleus content.
- RESTATEMENT: Satellite restates the nucleus in different words.
- EVALUATION: Satellite provides evaluation or assessment of the nucleus.
- INTERPRETATION: Satellite offers interpretation of the nucleus content.
- ATTRIBUTION: Satellite attributes information to a source.
- DEFINITION: Satellite defines a term or concept.
- CLASSIFICATION: Satellite classifies or categorizes information.

Figure 8: Relation Definitions for Intra-chunk RST Tree Construction.

Relation Definitions for Inter-chunk Rhetorical Graph Construction

Relation Definitions:

- SUPPORTS: Chunk provides support or evidence for another chunk.
- CONTRADICTS: Chunk contradicts or opposes another chunk.
- ELABORATES: Chunk elaborates on information in another chunk.
- EXEMPLIFIES: Chunk provides examples for another chunk's concepts.
- CAUSES: Chunk describes causes for events in another chunk.
- RESULTS_FROM: Chunk describes results from another chunk's events.
- ENABLES: Chunk describes what enables another chunk's situation.
- PREVENTS: Chunk describes what prevents another chunk's situation.
- PRECEDES: Chunk describes events that precede another chunk.
- FOLLOWS: Chunk describes events that follow another chunk.
- SIMULTANEOUS: Chunk describes simultaneous events with another chunk.
- BACKGROUND_FOR: Chunk provides background context for another chunk.
- GENERALIZES: Chunk provides general principles for another chunk's specifics.
- SPECIFIES: Chunk provides specific details for another chunk's generalizations.
- COMPARES_WITH: Chunk compares information with another chunk.
- CONTRASTS_WITH: Chunk contrasts information with another chunk.
- SUPPLEMENTS: Chunk supplements information in another chunk.
- REPLACES: Chunk replaces or updates information in another chunk.
- MOTIVATES: Chunk provides motivation for another chunk's content.
- JUSTIFIES: Chunk justifies claims or actions in another chunk.
- UNRELATED: Chunk has no meaningful rhetorical or semantic relation to another chunk.

Figure 9: Relation Definitions for Inter-chunk Rhetorical Graph Construction.

Prompt for Intra-chunk RST Tree Construction

You are an expert in Rhetorical Structure Theory (RST) analysis. Your task is to analyze the given text and construct a precise RST tree.

Critical instructions:

1. RST tree is a hierarchical tree structure (not a graph or network).
2. Each internal node has exactly two children: one nucleus (core) and one satellite (support) or two nuclei at the same time.
3. The nucleus contains the main information; the satellite provides supporting content.
4. Relations describe how the satellite relates to the nucleus.
5. Think carefully and output ONLY ONE complete RST tree.

Allowed RST relations:

ELABORATION, EXPLANATION, EVIDENCE, EXAMPLE, CONTRAST, COMPARISON, CONCESSION, ANTITHESIS, CAUSE, RESULT, CONSEQUENCE, PURPOSE, CONDITION, TEMPORAL, SEQUENCE, BACKGROUND, CIRCUMSTANCE, SUMMARY, RESTATEMENT, EVALUATION, INTERPRETATION, ATTRIBUTION, DEFINITION, CLASSIFICATION

Relation definitions:

{Relation Definition}

Step-by-step process:

1. Segment text into meaningful elementary discourse unit (EDU).
2. Determine the most important EDU (this becomes the root nucleus).
3. For each other EDU, decide: Is it nucleus (core) or satellite (support)?
4. Assign one relation from the allowed list.
5. Build the binary tree bottom-up.

Required output format:

`EDUs:`
[1] <first EDU>
[2] <second EDU>
...
[N] <Nth EDU>
`RST ANALYSIS:`
RELATION(EDU$_i$, EDU$_j$): {`RELATION TYPE`}
...
`TREE STRUCTURE:`

```
ROOT[1-N]
|--- NUCLEUS[X] <EDU text> (N)
|--- SATELLITE[Y] <EDU text> (S): {RELATION TYPE}
```

Validation rules:

- Each EDU must be complete and meaningful.
- Relations must be chosen from the allowed list.
- Mark (N) for nucleus, (S) for satellite.
- Output exactly ONE complete tree.

`TEXT TO ANALYZE:` {chunk$_i$}

Figure 10: Prompt for Intra-chunk RST Tree Construction. The relation definitions are provided in Figure 8.

Prompt for Listwise Discourse Relation Inference

You are an expert in discourse analysis. Your task is to infer the rhetorical relations jointly among a list of retrieved text chunks. In each call to this prompt, you are given the entire set of chunks, and you must construct a directed rhetorical graph over all of them.

Task objective:

Given a list of chunks `CHUNK[1]`, `CHUNK[2]`, ..., `CHUNK[K]`, determine for every ordered pair of distinct chunks whether there exists a meaningful rhetorical relation from the source chunk `CHUNK[i]` to the target chunk `CHUNK[j]`. If a relation exists, assign a directed discourse label; otherwise, mark the pair as `UNRELATED`.

Relation direction:

For each ordered pair (i, j) with $i \neq j$, treat `CHUNK[i]` as the source and `CHUNK[j]` as the target. The relation type should reflect how the source chunk contributes rhetorically to the target.

Allowed relation types:

SUPPORTS, CONTRADICTS, ELABORATES, EXEMPLIFIES, CAUSES, RESULTS_FROM, ENABLES, PREVENTS, PRECEDES, FOLLOWS, SIMULTANEOUS, BACKGROUND_FOR, GENERALIZES, SPECIFIES, COMPARES_WITH, CONTRASTS_WITH, SUPPLEMENTS, REPLACES, MOTIVATES, JUSTIFIES, UNRELATED

Relation definitions:

{Relation Definition}

Step-by-step process:

1. Carefully read all chunks in the list and identify the main claim, fact, or event expressed in each one.
2. Reason about how each chunk relates to the others at the discourse level, taking into account global context across all chunks.
3. For every ordered pair of distinct indices (i, j), decide whether `CHUNK[i]` serves a discourse function relative to `CHUNK[j]`.
4. If a rhetorical link exists, assign exactly one relation type from the allowed list.

Required output format:

For each ordered pair (i, j) with $i \neq j$, output one line in the following format:

`CHUNK[i] -> CHUNK[j]: {RELATION_TYPE}`

List all such lines for all ordered pairs in a consistent order (e.g., sorted by i then j).

Validation rules:

- Use only the allowed relation types.
- Relation direction must be from `CHUNK[i]` to `CHUNK[j]`.
- Output exactly one relation type for every ordered pair with $i \neq j$.

`TEXT TO ANALYZE:`
`CHUNK[1]`: [first chunk]
`CHUNK[2]`: [second chunk]
...
`CHUNK[K]`: [K-th chunk]

Figure 11: Prompt for listwise discourse relation inference. The relation definitions are provided in Figure 9.

Prompt for Discourse-Driven Planning

You are an expert in discourse-aware text generation. Your task is to produce a discourse-aware plan — a natural language paragraph that outlines how the final answer should be organized.

Inputs:

1. The user query.
2. Retrieved text chunks.
3. Intra-chunk RST trees, capturing local rhetorical hierarchies.
4. Inter-chunk rhetorical graph, modeling cross-passage discourse flow.

Critical instructions:

1. The plan must be written as a continuous paragraph in natural language.
2. The plan should describe the intended organization of the final answer.
3. The plan must be dynamically adapted to the given user query and evidence.
4. Avoid reproducing the content of the chunks; only outline how they will be used.
5. Output exactly one complete rhetorical plan.

Required output format:

`PLAN:` <one paragraph in natural language that describes the planned organization of the answer>
`TEXT TO ANALYZE:` {query, chunks, RST trees, rhetorical graph}

Figure 12: Prompt for Discourse-Driven Planning.

Prompt for Full Context Generation

You are an expert in question answering and text generation. Your task is to answer the user query using the provided full document as context.

Inputs:

1. The user query.
2. The full document.

Critical instructions:

1. The answer must directly address the user query.
2. Use the full document as the only source of factual claims.
3. If the document does not support a claim, do not add it.
4. Write a coherent answer without copying long spans verbatim from the document.

Required output format:

`ANSWER` <one paragraph or multiple paragraphs in natural language>
`TEXT TO ANALYZE` {query, document}

Figure 13: Prompt for full context generation used in our baseline.

Prompt for Standard RAG

You are an expert in retrieval-augmented generation. Your task is to answer the user query using only the retrieved text chunks as evidence.

Inputs:

1. The user query.
2. Retrieved text chunks.

Critical instructions:

1. The answer must directly address the user query.
2. Use the retrieved chunks as the only source of factual claims.
3. If the retrieved chunks do not support a claim, do not add it.
4. Write a coherent answer without copying long spans verbatim from the chunks.

Required output format:

`ANSWER` <one paragraph or multiple paragraphs in natural language>
`TEXT TO ANALYZE` {query, chunks}

Figure 14: Prompt for standard RAG used in our baseline.

Prompt for Retrieve-and-Plan Baseline

You are an expert in retrieval-augmented generation. Your task is to answer the user query by first writing a short plan and then writing the final answer using only the retrieved text chunks as evidence.

Inputs:

1. The user query.
2. Retrieved text chunks.

Critical instructions:

1. Write the plan as a single continuous paragraph that outlines the structure of the answer.
2. The plan must be grounded in what is supported by the retrieved chunks.
3. The answer must directly address the user query and use the retrieved chunks as the only source of factual claims.
4. If the retrieved chunks do not support a claim, do not add it.
5. Write a coherent answer without copying long spans verbatim from the chunks.

Required output format:

`PLAN` <one paragraph plan>
`ANSWER` <one paragraph or multiple paragraphs in natural language>
`TEXT TO ANALYZE` {query, chunks}

Figure 15: Prompt for the retrieve-and-plan baseline used in our ablation study.

Prompt for Plan-and-Retrieve Baseline

You are an expert in retrieval-augmented generation. Your task is to support a plan-guided retrieval procedure and then answer the user query.

Stage 1

Given only the user query, write a short plan and a retrieval hint that summarizes what evidence should be retrieved.

Stage 2

After retrieving all text chunks using the retrieval hint, write the final answer using only the retrieved text chunks as evidence.

Inputs:

1. The user query.
2. Retrieved text chunks returned after plan-guided retrieval.

Critical instructions:

1. Write the plan as a single continuous paragraph that outlines the structure of the answer.
2. The retrieval hint must be a list of retrieval queries that helps retrieve evidence aligned with the plan.
3. The answer must directly address the user query and use the retrieved chunks as the only source of factual claims.
4. If the retrieved chunks do not support a claim, do not add it.
5. Write a coherent answer without copying long spans verbatim from the chunks.

Required output format:

`PLAN` <one paragraph plan>
`RETRIEVAL HINT` <a list of retrieval queries>
`ANSWER` <one paragraph or multiple paragraphs in natural language>
`TEXT TO ANALYZE` {query, chunks}

Figure 16: Prompt for the plan-and-retrieve baseline used in our ablation study.

Prompt for Shallow Discourse Marker Inference

You are an expert in discourse analysis. Your task is to infer explicit discourse markers jointly among a list of retrieved text chunks. In each call to this prompt, you are given the entire set of chunks, and you must output a marker decision for every ordered pair of distinct chunks.

Task objective:

Given a list of chunks `CHUNK[1]`, `CHUNK[2]`, ..., `CHUNK[K]`, determine for every ordered pair (i, j) with $i \neq j$ whether there exists an explicit discourse marker from a marker list that indicates a meaningful rhetorical connection from `CHUNK[i]` to `CHUNK[j]`. If no marker is supported, output `NONE`.

Discourse marker list:

however, *but*, *although*, *in contrast*, *therefore*, *because*, *as a result*, *meanwhile*, *moreover*, *furthermore*, *for example*, *for instance*, *in addition*

Critical instructions:

1. For each ordered pair (i, j) with $i \neq j$, treat `CHUNK[i]` as the source and `CHUNK[j]` as the target.
2. Consider only explicit connectives that are supported by the two chunks. Do not infer implicit relations.
3. Output exactly one marker from the marker list if a marker is applicable; otherwise output `NONE`.
4. Output a decision for every ordered pair with $i \neq j$.

Required output format:

For each ordered pair (i, j) with $i \neq j$, output one line in the following format:

```
CHUNK[i] -> CHUNK[j]: {MARKER}
TEXT TO ANALYZE:
CHUNK[1]: [first chunk]
CHUNK[2]: [second chunk]
...
CHUNK[K]: [K-th chunk]
```

Figure 17: Prompt for discourse marker inference used in the shallow discourse marker baseline.

Prompt for Discourse-Guided RAG

You are an expert in retrieval-augmented generation with discourse knowledge. Your task is to generate a coherent and faithful answer by leveraging the following inputs:

Inputs:

1. The user query.
2. Retrieved text chunks.
3. Intra-chunk RST trees, capturing local rhetorical hierarchies.
4. Inter-chunk rhetorical graph, modeling cross-passage discourse flow.
5. A discourse-aware plan that outlines the intended argumentative organization.

Critical instructions:

1. The answer must directly address the user query.
2. Integrate evidence from multiple chunks, guided by their RST trees and rhetorical graph.
3. Follow the discourse-aware plan for structuring the answer.
4. Maintain factual accuracy, logical coherence, and rhetorical clarity.
5. Output a continuous answer in natural language.

Required output format:

`ANSWER:` <a single coherent paragraph or multi-paragraph answer grounded in discourse structures>

Validation requirements:

- The answer must be faithful to the retrieved content.
- The answer must be logically organized and reflect discourse-level coherence.
- Avoid verbatim repetition of chunks; instead, synthesize and integrate them.
- Output exactly one complete answer.

`TEXT TO ANALYZE:` {query, chunks, RST trees, rhetorical graph, discourse-aware plan}

Figure 18: Prompt for Discourse-Guided RAG.

Human Evaluation Guidelines

Prerequisites: Eligibility for this evaluation requires simultaneous fulfillment of two conditions: being a Master's or Ph.D. student in Computer Science or a closely related field, and demonstrating advanced proficiency in English sufficient to read and assess scientific news articles. Participants are compensated at the standard hourly rate and are asked to confirm that they meet these criteria before taking part in the task.

Instructions: For each selected sample, annotators are given the source document together with four anonymized summaries, and the system identities are hidden and the order is randomized for every instance. Raters are instructed to first read the source document carefully and then evaluate each summary independently using a three-point Likert scale along four criteria: *Relevance*, *Simplicity*, *Conciseness*, and *Faithfulness*.

Evaluation Criteria: Below, we provide a detailed explanation of the four criteria used in our human evaluation. Raters are asked to consider each criterion separately and to base their scores only on the information that is explicitly supported by the source document.

- **Relevance** This criterion assesses how well the summary covers the main topics, events, and findings discussed in the source document. A highly relevant summary focuses on central points and avoids spending space on marginal or tangential details.
- **Simplicity** This criterion measures how easy the summary is to read and understand. A simple summary uses clear and precise language, maintains a coherent structure, and avoids unnecessary jargon or convoluted phrasing that could hinder comprehension.
- **Conciseness** This criterion evaluates whether the summary is compact while still conveying the essential content. A concise summary avoids repetition and digression, omits minor details that are not needed for understanding, and does not exceed the length required to communicate the core message.
- **Faithfulness** This criterion judges whether the summary is supported by the source document and free of hallucinations. A faithful summary does not introduce claims that contradict the source, does not exaggerate or overgeneralize findings, and does not omit critical qualifications that change the meaning of the original text.

Rating System: For each criterion, raters assign an integer score from 1 to 3, where 1 indicates low quality, 2 indicates acceptable quality, and 3 indicates high quality. Scores should be given solely based on the source document and the summary, without using AI tools to assist in judgment. Annotators may consult trusted external resources, such as textbooks or scientific encyclopedias, only when they need to clarify terminology.

Figure 19: Guidelines presented to human raters for the SciNews dataset evaluation.