

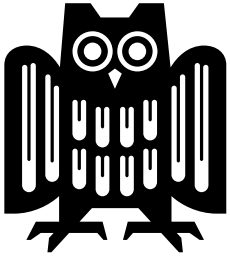
UNIVERSITÄT
DES
SAARLANDES

Discourse-Grounded Text Generation

Dongqi Liu

Doctor of Philosophy
Department of Language Science and Technology
Faculty of Philosophy
Saarland University

Saarbrücken, June 2026



UNIVERSITÄT
DES
SAARLANDES

UNIVERSITÄT DES SAARLANDES
PHILOSOPHISCHE FAKULTÄT

Discourse-Grounded Text Generation

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Philosophie
der Philosophischen Fakultät
der Universität des Saarlandes

vorgelegt von

Dongqi Liu

aus Hainan Province, People's Republic of China

Saarbrücken, 2026

Die Dekanin: Univ.-Prof. Dr. Nine Miedema
Dean of the Faculty of Humanities, Saarland University

Berichterstatter/innen: Prof. Dr. Vera Demberg
Full Professor, Saarland University

Prof. Dr. Frank Keller
Full Professor, University of Edinburgh

Prof. Dr. Junyi Jessy Li
Associate Professor, The University of Texas at Austin

Tag der letzten Prüfungsleistung: 29. Mai 2026

To

my father, Shichang Liu,
and my mother, Junying Li

*Language generation that neglects structure is like sand
in the desert wind, scattered and shapeless. With
structure, it becomes like trees in a forest, rooted,
branching, and richly layered.*

— *The Author*

Acknowledgements

I would first like to express my gratitude to my supervisor, Vera Demberg, for her guidance and support throughout my doctoral studies. Vera provided me with a great deal of intellectual freedom while consistently offering concrete and thoughtful feedback. It was under her encouragement that my reflections on how discourse structure could be made operational in summarization systems gradually took shape and eventually crystallized into the central thread of this dissertation. Completing this work under her supervision has been a truly rewarding experience.

I am also grateful to Mirella Lapata, who guided much of my research during my visiting period at the University of Edinburgh. With her clarity of thought and high standards, Mirella had a profound influence on the second part of my PhD, particularly on my work on multimodal summarization and discourse-driven planning-based generation. The collaboration within the Edinburgh NLP community broadened my understanding of what it means to conduct rigorous and impactful research, and I benefited greatly from its vibrant and open academic environment. In addition, I would like to thank Khalil Sima'an from the Institute for Logic, Language and Computation (ILLC) at the University of Amsterdam, whose supervision during my master's studies first sparked my interest in the intersection of structured representations and natural language generation. The seeds planted during that period continued to grow throughout my doctoral research.

I would also like to thank the members of my doctoral committee for accepting my invitation and for the time and expertise they devoted to the evaluation process. Vera Demberg, Frank Keller, and Junyi Jessy Li served as reviewers, Alexander Koller acted as the committee chair, and Josef van Genabith and Frances Yung served as committee members. Their insightful questions and detailed feedback helped improve the final version of this dissertation and made the defense itself a genuinely enriching academic exchange.

Much of the research presented in this dissertation was conducted in close collaboration with colleagues. I would like to thank Jia Loy, Louis Mahon, Yifu Qiu, Rohit Saxena, Xi Yu, Chenxi Whitehouse, Yifan Wang, and Zheng Zhao (listed in alphabetical order by surname) for their contributions at different stages of this dissertation. Coming from diverse countries and research backgrounds (i.e., China, the United Kingdom, Singapore, and India), each collaborator brought perspectives and ideas that would have been difficult to reach in isolation. These cross-institutional and cross-time-zone collaborations have been among the most valuable and cherished experiences of my doctoral journey.

All work carried out during my Ph.D. studies, including the research presented in this dissertation, is funded by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878). I appreciate this support, which provided the resources and freedom necessary to pursue my research. It also enabled me to explore ideas and conduct my work in an open and collaborative environment.

I would like to express my deepest gratitude to my parents, Mr. Shichang Liu and Mrs. Junying Li, for their unwavering support and trust throughout my growth and education. Although they could not take part in the technical discussions of my research, they always believed in the path I chose. That steadfast trust, more than any concrete assistance, has been a constant source of reassurance. This dissertation is dedicated to my parents.

Finally, I would also like to thank myself. Much of this journey was lived quietly and, in many ways, alone in Germany, far from home and from the people who knew me best. There were difficult moments that only I fully understood, and some that I still prefer not to put into words. Yet I continued, step by step, through exhaustion and doubt. Finishing this dissertation does not erase those moments, but it gives them shape and meaning. For that quiet persistence, I am deeply grateful.



European Research Council
Established by the European Commission

Abstract

This dissertation investigates discourse-level challenges in text generation tasks where the input is lengthy, evidence is structurally fragmented, or relevant information needs to be retrieved from external sources. While mainstream neural generation systems tend to produce fluent outputs at the sentence level, they frequently encounter difficulties in maintaining balanced content selection, preserving inter-paragraph coherence, and sustaining a logical argumentative flow across longer spans. To address these limitations, this dissertation conceptualizes discourse structure as a source of prior knowledge that can be explicitly modeled and effectively utilized. This dissertation explores a progressive trajectory that spans the development of structural representations and their integration methods, the transition from full fine-tuning to parameter-efficient adaptation, the shift from purely textual input to multimodal evidence, the extension from expert-written summaries to layperson-accessible explanations, and the application of discourse modeling to retrieval-augmented generation where retrieved evidence lacks explicit rhetorical connections. This trajectory aims to enhance the operability and controllability of discourse signals across summarization and retrieval-augmented generation systems.

In long-document summarization settings, the dissertation adopts Rhetorical Structure Theory (RST) as the principal discourse framework, emphasizing the importance of two key aspects: relation types and structural uncertainty. On the one hand, different discourse relations imply distinct preferences regarding what to retain or omit; relying solely on nucleus–satellite centrality while ignoring relation types fails to capture the semantic rationale behind inter-sentential connections. On the other hand, automatic discourse parsing is prone to errors in out-of-domain settings, and injecting a single parse tree as a hard constraint can amplify these errors during training and entrench them in the model’s generation behavior. To mitigate this, the dissertation proposes distributed representations that preserve parsing uncertainty by transforming discrete tree structures into probabilistic tensors or matrices aligned with neural computation. These soft sig-

nals are then injected into attention mechanisms or parameter update paths as flexible guidance.

As the form of evidence expands from text to multimodal inputs, end-to-end generation becomes incapable of reproducing the characteristic discourse organization of target summary genres. To address this issue, this dissertation constructs and releases a large-scale aligned dataset for scientific presentation videos and benchmarks multiple multimodal large models under varying fine-tuning settings. This dissertation further introduces planning as an explicit intermediate variable, representing the latent structure of summaries through ordered sequences of questions. In the context of layperson-oriented summarization, this dissertation incorporates explanatory content as a controllable objective. This dissertation identifies explanatory units and their targets via discourse parsing and transforms implicit questions into explicit plans to guide the emergence and functional placement of explanations. To address the evaluation challenge posed by helpful external additions being misclassified as hallucinations, this dissertation proposes an improved consistency metric based on external knowledge verification, complemented by human evaluation. Beyond the summarization settings explored above, this dissertation extends discourse structure modeling to retrieval-augmented generation (RAG). The proposed framework constructs local RST trees within each retrieved chunk to capture intra-chunk discourse hierarchy and infers a directed rhetorical graph across chunks to model inter-chunk relations such as support, contrast, and elaboration. A discourse-driven planning module then synthesizes these structural signals into an ordered blueprint that guides the final generation.

In summary, this dissertation presents a coherent path from distributed discourse signals to plan-based intermediate structures, and extends these principles to the organization of retrieved evidence in retrieval-augmented generation. Under various constraints, including long inputs, limited fine-tuning capacity, multimodal evidence, and structurally disconnected retrieved passages, explicit modeling of organizational decisions consistently enhances controllability and factual reliability. This work contributes reusable data resources, modeling strategies, and evaluation frameworks that support future research on discourse-guided text generation.

Keywords: Long Input Summarization; Retrieval-Augmented Generation; Rhetorical Structure Theory; Question under Discussion.

Zusammenfassung

Diese Doktorarbeit untersucht diskursbezogene Herausforderungen in Textgenerierungsaufgaben, bei denen die Eingabe sehr lang ist, die Evidenz strukturell fragmentiert vorliegt oder relevante Informationen zunächst aus externen Quellen abgerufen werden müssen. Während gängige neuronale Generierungssysteme auf Satzebene meist flüssige Ausgaben erzeugen, haben sie häufig Schwierigkeiten, eine ausgewogene Inhaltsauswahl sicherzustellen, die Kohärenz zwischen Absätzen zu bewahren und über längere Textspannen hinweg einen logisch konsistenten argumentativen Textverlauf aufrechtzuerhalten. Um diesen Einschränkungen zu begegnen, konzeptualisiert diese Doktorarbeit Diskursstruktur als eine Quelle von Vorwissen, die explizit modelliert und gezielt nutzbar gemacht werden kann. Sie zeichnet eine Entwicklung nach, die vom Aufbau struktureller Repräsentationen und ihrer Integration in Modelle über den Übergang vom vollständigen Fine-Tuning zur parametereffizienten Anpassung, den Wechsel von rein textbasierten zu multimodalen Eingaben sowie die Erweiterung von expert:innenverfassten Zusammenfassungen hin zu für Laien zugänglichen Erklärungen bis zur Anwendung der Diskursmodellierung auf Retrieval-gestützte Generierung reicht, bei der identifizierte Quellen keine expliziten rhetorischen Verknüpfungen aufweist. Diese Entwicklung zielt darauf ab, die Operationalisierbarkeit und Kontrollierbarkeit von Diskurssignalen in Zusammenfassungs- und Retrieval-gestützten Generierungssystemen zu verbessern.

Im Kontext der Zusammenfassung langer Dokumente verwendet die Doktorarbeit die “Rhetorical Structure Theory” (RST) als zentralen diskurstheoretischen Ansatz und betont die Bedeutung zweier Kernaspekte: der Relationstypen und der strukturellen Unsicherheit. Zum einen implizieren unterschiedliche Diskursrelationen unterschiedliche Präferenzen dafür, was in einer Zusammenfassung beibehalten oder weggelassen werden sollte; eine ausschließliche Orientierung an der Zentralität von “nucleus”- und “satellite”-Einheiten unter Ausblendung der Relationstypen erfasst die semantische Motivation intersentenzieller Verknüpfungen nicht hinreichend. Zum anderen ist die automatische

Diskursanalyse in fachfremden Anwendungsbereichen fehleranfällig, und das Einbringen eines einzelnen Analysebaums als harte Vorgabe kann diese Fehler während des Trainings verstärken und sie im Generationsverhalten des Modells verfestigen. Um dem entgegenzuwirken, schlägt die Doktorarbeit verteilte Repräsentationen vor, die die Unsicherheit der Diskursanalyse bewahren, indem diskrete Baumstrukturen in probabilistische Tensoren oder Matrizen überführt werden, die auf neuronale Berechnung abgestimmt sind. Diese weichen Signale werden anschließend als flexible Leitinformationen in Aufmerksamkeitsmechanismen oder in Pfade der Parameteraktualisierung eingebracht.

Mit der Ausweitung der zugrundeliegenden Daten von Text auf multimodale Eingaben sind end-to-end Generierungsansätze nicht mehr in der Lage, die charakteristische Diskursorganisation der Zielgattungen von Zusammenfassungen zuverlässig zu reproduzieren. Um dieses Problem zu adressieren, erstellt und veröffentlicht diese Doktorarbeit einen groß angelegten, zeitlich alignierten Datensatz für wissenschaftliche Vortragsvideos und vergleicht mehrere multimodale große Modelle unter unterschiedlichen Fine-Tuning-Einstellungen systematisch miteinander. Darüber hinaus führt die Doktorarbeit Planung als explizite Zwischenvariable ein, die die latente Struktur von Zusammenfassungen durch geordnete Folgen von Fragen repräsentiert. Im Kontext laienorientierter Zusammenfassungen integriert diese Doktorarbeit erklärende Inhalte als steuerbares Ziel. Sie identifiziert Erklärungseinheiten und ihre Zielobjekte mittels Diskursanalyse und überführt implizite Fragen in explizite Pläne, um das Entstehen und die funktionale Platzierung von Erklärungen gezielt zu steuern. Um die Evaluationsherausforderung zu adressieren, dass hilfreiche externe Ergänzungen fälschlich als Halluzinationen klassifiziert werden, schlägt diese Doktorarbeit eine verbesserte Konsistenzmetrik vor, die auf externer Wissensüberprüfung basiert und durch menschliche Evaluation ergänzt wird. Über die oben untersuchten Zusammenfassungsszenarien hinaus erweitert diese Doktorarbeit die Diskursstrukturmodellierung auf die Retrieval-gestützte Generierung (RAG). Der vorgeschlagene Ansatz konstruiert innerhalb jedes abgerufenen Textabschnitts lokale RST-Bäume zur Erfassung der intrachunkalen Diskurshierarchie und leitet über alle Textabschnitte hinweg einen gerichteten rhetorischen Graphen ab, der Inter-Chunk-Relationen wie Unterstützung, Kontrast und Elaboration modelliert. Ein diskursgesteuertes Planungsmodul verdichtet diese strukturellen Signale anschließend zu einem geordneten Bauplan, der die abschließende Generierung steuert.

Zusammenfassend präsentiert diese Doktorarbeit einen kohärenten Entwicklungs-

pfad von verteilten Diskurssignalen hin zu planbasierten Zwischenstrukturen und erweitert diese Prinzipien auf die Organisation abgerufener Evidenz in der Retrieval-gestützten Generierung. Unter verschiedenen Einschränkungen, darunter lange Eingaben, begrenzte Fine-Tuning-Kapazitäten, multimodale Evidenz und strukturell unverbundene abgerufene Passagen, erhöht die explizite Modellierung organisatorischer Entscheidungen durchgängig die Kontrollierbarkeit und die faktische Verlässlichkeit generierter Texte. Diese Arbeit leistet Beiträge in Form wiederverwendbarer Datenressourcen, Modellierungsstrategien und Evaluationsrahmen, die zukünftige Forschung zur diskursgeleiteten Textgenerierung unterstützen.

Schlüsselwörter: Zusammenfassung langer Eingaben; Retrieval-gestützte Generierung; Rhetorical Structure Theory; Question under Discussion.

Contents

1	Introduction	1
1.1	Research Background	1
1.2	Research Challenges	4
1.3	Research Questions	5
1.4	Dissertation Organization	11
1.5	Published Work	13
1.6	Statement of Individual Contributions	15
2	Background	17
2.1	Problem Setting and Task Scope	17
2.1.1	Evidence-Conditioned Generation	18
2.1.2	Model Paradigms for Text Generation	20
2.1.3	Long Inputs and Attention Patterns	21
2.2	Discourse Structure	22
2.2.1	Rhetorical Structure Theory	22
2.2.2	Question Under Discussion	24
2.3	Resource Constraints and Efficient Adaptation	26
2.3.1	PEFT Methods	26
2.3.2	Challenges of PEFT Methods	28
2.4	Retrieval-Augmented Generation	29
2.4.1	The Retrieve-then-Generate Paradigm	29
2.4.2	Challenges of RAG Methods	31
2.5	Planning and Structured Generation	32
2.5.1	Plans as Explicit Intermediate Targets	32
2.5.2	Planning across Modalities and Audiences	34
2.6	Evaluation and Verification	35

CONTENTS

2.6.1	Automatic Quality Metrics	36
2.6.2	Grounding and Factuality Metrics	37
2.6.3	Human Evaluation	38
2.6.4	LLM-as-a-Judge	39
2.7	Chapter Summary	40
3	Discourse-Aware Attention for Text Summarization	41
3.1	Problem Formulation	42
3.2	Methodology	44
3.3	Experimental Setup	48
3.4	Results and Analysis	50
3.5	Related Work	55
3.6	Chapter Summary	56
4	Discourse-Aware LoRA for Text Summarization	59
4.1	Problem Formulation	60
4.2	Methodology	62
4.3	Experimental Setup	65
4.4	Results and Analysis	67
4.5	Related Work	78
4.6	Chapter Summary	79
5	Discourse-Aware Planning for Video-to-Text Summarization	81
5.1	Problem Formulation	83
5.2	The VISTA Dataset	85
5.3	Benchmarking VISTA	88
5.4	Experimental Setup	90
5.5	Results and Analysis	92
5.6	Related Work	108
5.7	Chapter Summary	109
6	Discourse-Aware Planning for Explanatory Summarization	111
6.1	Problem Formulation	112
6.2	Methodology	115
6.3	Experimental Setup	118

CONTENTS

6.4	Results and Analysis	120
6.5	Related Work	135
6.6	Chapter Summary	137
7	Discourse-Aware Retrieval-Augmented Generation	139
7.1	Problem Formulation	140
7.2	Methodology	143
7.3	Experimental Setup	145
7.4	Results and Analysis	146
7.5	Related Work	161
7.6	Chapter Summary	162
8	Conclusion	163
8.1	Core Contributions	164
8.2	Synthesis of Findings	165
8.3	Research Limitations	170
8.4	Future Directions	171
8.5	The Use of AI Tools	174
8.6	Ethical Considerations	174
Appendices		
A	Appendix for Chapter III	177
A.1	Questionnaire of Human Evaluation	177
B	Appendix for Chapter IV	179
B.1	GPT-4 Prompts	179
B.2	Human Evaluation Guidelines	179
C	Appendix for Chapter V	183
C.1	Data Sample	183
C.2	Prompts Used in Our Study	183
C.3	Human Evaluation Guidelines	183
C.4	Prompt for GPT-o1 to Evaluate Summary Quality	183

CONTENTS

D Appendix for Chapter VI	189
D.1 Rule-based Extraction of Explanations	189
D.2 Human Expert Evaluation Results	192
D.3 Model Prompts	192
D.4 Human Evaluation Guidelines	192
E Appendix for Chapter VII	197
E.1 Details of Evaluation Metrics	197
E.2 Details of Baselines	199
E.3 Prompts for Disco-RAG	200
E.4 Human Evaluation Guidelines	203
List of Figures	212
List of Tables	217
Bibliography	221

1

Introduction

This chapter introduces the motivation, research questions, and overall scope of the dissertation. The central argument is that discourse structure, when explicitly modeled and integrated into the generation process, can serve as a reliable organizing principle for text generation tasks involving long, fragmented, or externally retrieved evidence. [Section 1.1](#) situates this argument within the broader context of automatic summarization and retrieval-augmented generation. [Section 1.2](#) identifies the key limitations of existing approaches that motivate the explicit incorporation of discourse-level signals. Building on these observations, [Section 1.3](#) formulates five research questions and their associated hypotheses, tracing a progressive trajectory from discourse representation to planning-based generation. [Section 1.4](#) outlines the structure of the dissertation and describes how the individual chapters contribute to the overarching narrative. Finally, [Section 1.5](#) lists the publications on which the core chapters are based, [Section 1.6](#) specifies the individual contributions of the doctoral candidate.

1.1 Research Background

Efficient access to information is a longstanding concern across academic research and industrial practice, and has driven sustained interest in systems that can synthesize complex evidence into coherent and faithful textual outputs ([Pagnoni et al. 2023](#); [Dahan et al. 2025](#); [Cao et al. 2024](#)). A single research topic may produce hundreds of publications annually, while organizational workflows continuously accumulate large volumes of internal records, such as business reports, technical specifications, and system logs

(Pratapa et al. 2023; Overbay et al. 2023; Wang et al. 2023a). Under these conditions, two complementary needs have become prominent. On the one hand, automatic summarization helps readers navigate long documents by compressing their content into concise representations that foreground main claims and supporting evidence (Takeshita et al. 2024; Pagnoni et al. 2023; Hua et al. 2024). On the other hand, retrieval-augmented generation (RAG) addresses knowledge-intensive scenarios where the relevant evidence should first be gathered from external corpora before a coherent response can be produced (Lewis et al. 2020b; Asai et al. 2024). Despite their different interfaces, both tasks face a shared underlying challenge: The generation system needs to identify salient propositions, understand how distributed pieces of evidence relate to one another, and organize them into outputs that faithfully reflect the source reasoning.

In academic writing, long documents rarely consist of independent facts; instead, they articulate an extended argument that connects research motivation, methodological decisions, and empirical findings (Liu et al. 2023c; Liu et al. 2024a; Takeshita et al. 2024). Interpreting a particular claim often requires tracing dependencies that span multiple sections of the document (Wang et al. 2023b; Zhong et al. 2025b). A similar pattern is evident in technical reports, where the justification for a recommendation may draw on risk analyses, design constraints, or implementation considerations that are introduced earlier or elaborated later than the point at which the final decision is stated. When evidence is retrieved from external sources rather than drawn from a single document, these dependency structures become even more difficult to reconstruct, as the retrieved passages typically lack the connective tissue that would link them into a coherent argument. Therefore, an effective generation system should assist in navigating these distributed relationships and in reconstructing the chain of reasoning that underlies the central conclusions (Liu et al. 2023c; Syed et al. 2023; Gurjar et al. 2025).

As large language models (LLMs) are increasingly applied to text generation systems, the limitations of strategies that rely primarily on surface-level cues (for example, the autoregressive mechanism for next-token or next-sentence prediction) have become more noticeable (Zhong et al. 2025b). The generated outputs may appear coherent and well-organized, yet still overlook essential qualifications from the source text or merge distinct pieces of evidence into arguments that exceed the evidence delivered by the original material (Tang et al. 2023b; Jeong et al. 2025). These deficiencies are particularly pronounced in long-document scenarios (Gao et al. 2025), where architectural constraints such as limited attention windows and position-dependent degradation

restrict reliable access to information distributed across distant segments of the input (Peng et al. 2024; Song et al. 2024a). In retrieval-augmented settings, the difficulties are compounded by the absence of explicit discourse connections among independently retrieved passages (Edge et al. 2024), leaving the generator without a rhetorical scaffold for evidence integration. As a result, decisions about content selection and structural organization are, in both settings, more likely to be steered by local lexical or positional signals than by the underlying argumentative structure of the evidence.

However, surface-level similarity alone cannot provide adequate guidance for generating outputs that satisfy both structural coherence and faithful alignment with the source reasoning (Mondshine et al. 2025). High-quality generation from long or structurally fragmented inputs requires modeling signals that extend beyond lexical overlap or locally salient patterns (Liu et al. 2023c; Liu et al. 2024a). In other words, effective models need to identify central propositions, distinguish main claims from supporting or contextual content, and preserve the discourse relations that govern how the overall argument is constructed (Gao et al. 2023). These considerations emphasize the importance of incorporating discourse-level information and modeling strategies that provide explicit structural support for controlled content organization and reliable interpretation (Pagnoni et al. 2023).

From a discourse perspective, a document is not merely an ordered sequence of sentences but a collection of semantically coherent units connected by rhetorical relations that jointly structure meaning and argumentation (Liu et al. 2025c; Miao et al. 2025; Eichin et al. 2025). Rhetorical Structure Theory (RST) offers a well-established framework for modeling this organization by distinguishing units that contribute directly to a document’s central argumentative content from those that provide background information, clarification, or supplementary detail (Liu et al. 2023c; Liu et al. 2024a). This distinction is also related to the Question under Discussion (QUD) framework, which conceptualizes discourse as organized around a hierarchy of implicit questions that determine which information is relevant at a given point. Within the QUD perspective, central units address the active question, whereas peripheral units refine, constrain, or contextualize the space of possible answers. Consistent with these theoretical accounts, empirical studies of human reading behavior demonstrate that human readers tend to follow the progression of the main argument, evaluate supporting evidence, attend to assumptions that condition the claims, and register constraints that delimit the scope of the conclusions (Shubi et al. 2025; Miao et al. 2025).

1.2 Research Challenges

Building on the foregoing discussion, the central claim of this dissertation is that explicitly modeling discourse structure provides a means of alleviating the limitations of surface-level generation models. By incorporating discourse knowledge into the generation process, models can be guided toward content selection and organization decisions that better reflect the argumentative structure of the source input.

Although a growing body of work provides both theoretical motivation and empirical evidence for the benefits of discourse-aware summarization/generation (Gerani et al. 2014; Kikuchi et al. 2014; Cohan et al. 2018; Xu et al. 2020; Dong et al. 2021), current approaches often fall short of systematically and robustly leveraging discourse structure within modern neural architectures. The core challenge is the difficulty of operationalizing discourse information in ways that are both architecturally compatible and empirically effective (Miao et al. 2025). These challenges include representing discourse information that is often noisy in a form that can be reliably learned; introducing discourse guidance under constrained adaptation budgets with limited trainable parameters; leveraging discourse cues in multimodal summarization to improve faithfulness and controllability; employing discourse structure to balance readability and reliability when generating explanatory summaries for non-expert audiences; and organizing structurally disconnected retrieved evidence through discourse relations to support more factually grounded generation in retrieval-augmented settings. Taken together, these challenges delineate a set of structural and methodological issues that are examined in detail in the following discussion.

- Existing discourse-aware summarization approaches typically rely on a single discourse parse generated by an automated parser and treat this representation as fixed throughout the training and inference (Gerani et al. 2014; Kikuchi et al. 2014; Cohan et al. 2018; Xu et al. 2020). However, this design choice restrains the effective use of relation-type semantics and fails to account for the uncertainty and variability inherent in automatically induced discourse structures.
- Existing parameter-efficient fine-tuning techniques are designed to adapt large pretrained models by updating a small number of parameters (Li et al. 2021; Lester et al. 2021; Hu et al. 2022; Liu et al. 2022c). However, available empirical evidence suggests that these techniques struggle to reliably encode discourse-level

information, including content salience, evidence ordering, and rhetorical emphasis.

- Existing video-to-text summarization systems largely emphasize aligning visual content with corresponding textual descriptions (Palaskar et al. 2019; Li et al. 2017; Liu et al. 2021a; Kudo et al. 2023). However, the absence of an explicit discourse-guided planning component restricts their capability to organize information in a manner that reflects the rhetorical conventions of abstracts.
- Existing public-facing summarization methods generate fluent outputs with limited mechanisms for controlling the introduction and placement of explanatory content (Zhang et al. 2024c; Joseph et al. 2024; Cardenas et al. 2023). However, the presence and ordering of explanations can substantially influence how readers assess the reliability and trustworthiness of the generated lay summaries.
- Existing retrieval-augmented generation systems concatenate retrieved passages as flat context without modeling the rhetorical relations within or across them (Lewis et al. 2020b; Edge et al. 2024). However, when evidence is distributed across multiple chunks that may contain conflicting, complementary, or conditionally scoped claims, the absence of explicit discourse structure can lead to overgeneralization, factual inconsistency, or incoherent reasoning in the generated output.

1.3 Research Questions

In response to the research challenges outlined above, this dissertation treats discourse modeling not as a monolithic solution but as a group of interconnected research components operating under specific constraints. Accordingly, the dissertation formulates a core research question: Discourse structure can function as a valuable signal for improving long-form text generation. This overarching inquiry is decomposed into a series of specific research questions, each paired with a corresponding hypothesis and associated with a distinct modeling challenge. The progression of these questions traces a trajectory from foundational representation and integration of discourse signals, to their behavior under constrained adaptation, to their utility in multimodal summarization settings,

to their role in enhancing explanation quality and factual reliability in public-facing contexts, and finally to their application in organizing retrieved evidence for retrieval-augmented generation. The specific research questions addressed in this dissertation are listed below.

Research Question 1: *How can discourse structure be transformed into uncertainty-aware signals that support content selection and ordering in long-document summarization?*

This research question investigates the challenge of operationalizing discourse structure as a learnable signal in long-document summarization, under the realistic assumption that discourse annotations obtained from automatic parsers are noisy and uncertain. Rather than treating discourse analysis as a fixed preprocessing step that produces a single deterministic structure, the focus is on how uncertainty in discourse relations and centrality assignments can be explicitly modeled and propagated into the summarization process.

Concretely, this question examines whether distributions over discourse relations and nucleus–satellite assignments can be integrated into neural summarization models in a way that directly influences content selection and sentence ordering. The core intuition is that preserving alternative discourse analyses, instead of committing to a single hard label, allows the model to exploit discourse information when it is informative while remaining robust to parser errors when it is not. This perspective reframes discourse structure from a brittle constraint into a soft organizing signal that can be weighed against other evidence during generation.

Hypothesis I: Explicitly representing discourse structure as a distribution can mitigate information loss caused by incorrect or overconfident discourse labels, enabling summarization models to account for uncertainty and make more robust content selection and ordering decisions.

This hypothesis is investigated in [Chapter 3](#), which proposes a discourse-guided summarization framework that injects probabilistic discourse signals into the attention mechanisms of Transformer-based models. By modulating attention using distributions over rhetorical relations, the model is encouraged to prioritize discourse-salient content while remaining flexible under structural ambiguity. The empirical analysis examines how different relation types affect salience modeling, how parser uncertainty propagates

through the model during training and inference, and how these design choices impact summary quality.

Evaluation is conducted along multiple dimensions, including content coverage, discourse coherence, and factual consistency, to ensure that improvements in organization do not come at the expense of faithfulness or reliability. This investigation establishes the conditions under which discourse structure can function as a reliable and actionable organizing signal in long-input summarization, providing the methodological foundation for the subsequent research questions.

Research Question 2: *How can discourse signals reliably guide content selection and organization under parameter-efficient adaptation regimes with limited update capacity and computational resources?*

Building on the investigation of how RST structures can serve as uncertainty-aware signals (Research Question 1), this research question explores their effectiveness under constrained model adaptation. In parameter-efficient fine-tuning (PEFT) settings such as low-rank adaptation (LoRA), only a small number of parameters are updated, restricting the model’s capacity to revise internal representations. While PEFT methods often improve surface-level fluency, they often fall short in shaping global discourse structure. This setting thus offers a rigorous test of whether discourse-level signals can still influence document organization under strict memory and computing constraints.

Hypothesis II: When integrated as structured and uncertainty-aware guidance, discourse signals can exert an influence on content selection and inter-sentential organization under constrained parameter update budgets.

This hypothesis is examined in [Chapter 4](#), which extends the discourse-guided framework to PEFT settings by introducing low-rank variants of discourse integration. In particular, the model injects uncertainty-aware discourse signals into a frozen backbone using compact, learnable modules, allowing structural information to influence generation without incurring the high memory costs of full fine-tuning. This chapter further explores how these signals interact with different adaptation configurations and evaluates their effectiveness across salience-sensitive summarization tasks. Special attention is given to the stability of discourse behavior and its impact on factual consistency when adaptation capacity is limited.

Evaluation is conducted under both quality and efficiency metrics, including summary structure, coherence, factuality, and training resource usage, to determine whether

improvements in discourse control are both measurable and scalable. The results provide insight into the conditions under which discourse-level guidance remains viable even when the model’s plasticity is severely constrained.

Research Question 3: *How can discourse-aware planning be extended to support macro-level organization and factual grounding in video-to-text summarization, given the challenges of fragmented and multimodal input?*

Building on the previous investigations into discourse guidance under textual and adaptation constraints (Research Questions 1 and 2), this research question extends the scope to multimodal inputs, where evidence is distributed across text, audio, and visuals. Scientific presentation videos often convey information in a temporally fragmented and structurally diffuse manner, requiring models not only to extract salient content but also to recombine it into a coherent and factually grounded textual abstraction. This setting raises the question of whether an explicit, discourse-informed planning mechanism can help bridge the gap between low-level modality streams and high-level structured summaries.

Hypothesis III: When cross-modal alignment is sufficiently reliable, discourse-aware planning enables summarization models to produce more coherent and factually grounded outputs by structuring generation around an abstract representation of argumentative intent.

This hypothesis is investigated in [Chapter 5](#), which introduces a new multimodal summarization framework tailored for scientific video-to-text generation. This chapter begins by constructing a large-scale dataset that pairs academic presentation videos with their corresponding paper abstracts. The core modeling strategy is a two-stage generation pipeline. In the first stage, the model generates a discourse-aware plan, a structured outline of rhetorical moves or high-level intents inferred from video content. In the second stage, the model conditions on this plan to produce a full-length summary, allowing it to prioritize discourse-salient segments and maintain coherence over temporally dispersed input signals. The planning mechanism is informed by automatically derived rhetorical cues and reflects the communicative intent often present in structured presentations.

The evaluation is conducted along four dimensions, including structural coherence, cross-modal alignment, content coverage, and factual accuracy. Special attention is

given to the interaction between planning structure and modality noise—for example, how visual gaps or speech hesitations influence the downstream generation. The empirical results demonstrate that discourse-informed planning improves both organizational structure and grounding fidelity, particularly in complex or sparsely aligned video segments.

Research Question 4: *How can explanatory content in public-facing summarization be treated as a controllable discourse element to improve accessibility and readability without compromising factual reliability?*

Extending the prior investigation of discourse-aware planning in multimodal settings (Research Question 3), this research question shifts the focus to lay summarization, where models should balance informativeness with accessibility for non-expert audiences. In such settings, explanatory content plays a crucial role in supporting comprehension, but also introduces risk, particularly when models rely on background knowledge or extrapolate beyond source material. This raises the challenge of structuring explanations in a principled and controllable way, such that they enhance readability without undermining factual fidelity.

Hypothesis IV: When modeled as controllable discourse elements, explanations can enhance the accessibility and readability of public-facing summaries while preserving factual reliability, provided that their discourse scope, target specificity, and placement are explicitly managed.

This hypothesis is investigated in [Chapter 6](#), which introduces an explanation-sensitive summarization framework grounded in discourse planning. The model operates in a two-stage fashion. It first generates a high-level rhetorical plan that explicitly marks where and how explanatory content should appear (e.g., as background, elaboration, analogy), and then conditions generation on this plan to ensure rhetorical coherence and factual alignment. This design treats explanation as a planning unit, rather than an emergent side effect, and enables control over its presence, function, and position within the output.

To ensure reliability, the framework further incorporates constraints that assess the factual risk associated with each explanatory unit. These constraints help prioritize factual anchoring and de-emphasize hallucination-prone elaborations. The evaluation protocol combines automatic and human evaluations to jointly assess explanatory utility, accessibility, and factual integrity, including specific checks for unverifiable or audience-misleading content. Special attention is paid to how target-audience specificity (e.g.,

expert vs lay) interacts with explanation behavior and perceived quality. The findings demonstrate that explanation can be rendered more reliable and audience-appropriate when framed as a discourse-guided generation problem, rather than left to ad hoc decoding dynamics.

Research Question 5: *How can discourse structure be leveraged to organize structurally disconnected retrieved evidence, so that retrieval-augmented generation produces more coherent and factually consistent outputs?*

The preceding four research questions investigate discourse modeling within settings where all relevant evidence is contained within a given source input. In retrieval-augmented generation (RAG), this assumption no longer holds. The evidence needed to produce a faithful response is usually first gathered from an external corpus, and the resulting chunks lack the discourse continuity present in coherent documents. Standard RAG pipelines offer no mechanism for modeling rhetorical relations within or across retrieved passages. This research question asks whether the discourse modeling methods developed in this dissertation, namely RST-based structural analysis and QUD-inspired planning, can be adapted to construct a rhetorical scaffold over retrieved evidence that supports evidence integration and answer organization.

Hypothesis V: Explicitly constructing intra-chunk discourse trees and inter-chunk rhetorical graphs over retrieved evidence, combined with discourse-driven planning, can improve the factual consistency, coherence, and answer quality of retrieval-augmented generation.

This hypothesis is investigated in [Chapter 7](#), which proposes Disco-RAG, a framework that augments standard RAG with three discourse-aware components. First, an RST tree is constructed for each retrieved chunk to capture its local discourse hierarchy. Second, a directed rhetorical graph is inferred across all chunks to model inter-chunk relations such as support, contrast, and elaboration. Third, a planning module synthesizes these structural signals into an ordered blueprint that guides the final generation.

The evaluation examines whether this combination of local, global, and planning-level discourse modeling yields measurable improvements in answer quality and factual grounding across knowledge-intensive QA and summarization benchmarks. Ablation studies isolate the contribution of each component, and controlled perturbation experiments assess the sensitivity of the framework to structural noise in the parsed discourse

representations, providing evidence that the observed gains stem from the quality of discourse modeling rather than from the mere presence of additional structural tokens.

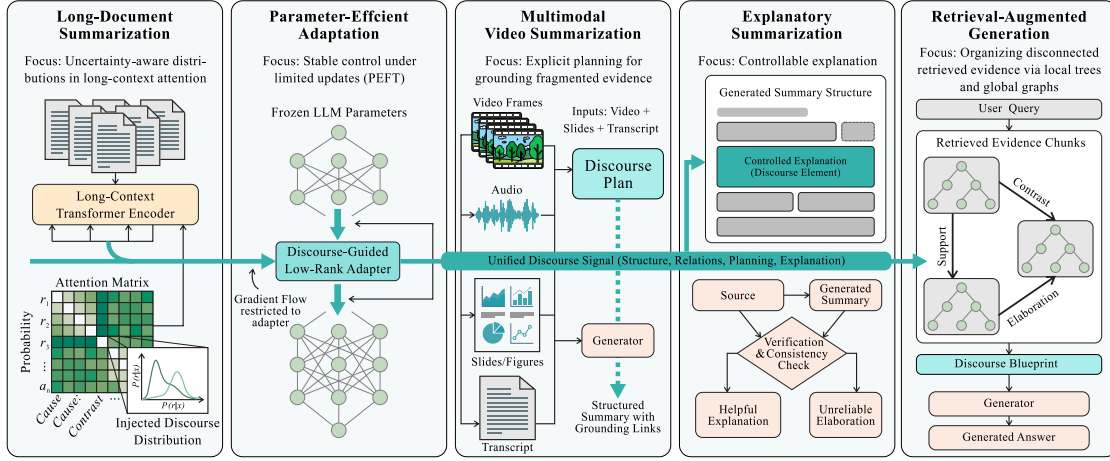


Figure 1.1: Dissertation roadmap showing how discourse signals are incorporated across the five research questions.

Figure 1.1 summarizes the scope and progression of the five research questions. Together, they trace a trajectory from uncertainty-aware discourse modeling (*RQ1*), through adaptation-constrained learning (*RQ2*), to multimodal application (*RQ3*), to explanation-sensitive generation (*RQ4*), and to discourse-aware retrieval-augmented generation (*RQ5*). Each research question is grounded in a corresponding hypothesis and targets a distinct set of technical conditions, while collectively advancing a shared objective: Establishing discourse as a reliable and empirically testable organizing principle for long-form text generation.

1.4 Dissertation Organization

The research questions outlined in the preceding section delineate five interrelated challenges that arise from limitations in existing research: representing discourse as an uncertainty-aware signal; integrating such signals under constrained adaptation settings; extending the discourse-based framework to multimodal scientific summarization; modeling explanatory content in public-facing summaries; and organizing retrieved evidence through discourse structure to support retrieval-augmented generation. Addressing these challenges requires methodological contributions that span input modalities, modeling constraints, task interfaces, and evaluation regimes. To support this goal, this disserta-

tion is organized into eight chapters, with **Chapter 3** through **Chapter 7** devoted to the technical contributions. These five technical chapters build on one another, progressively developing and empirically examining the proposed discourse-guided framework across abstractive summarization and retrieval-augmented generation settings.

Chapter 1 introduces the motivation of the dissertation by situating text generation as a task constrained not only by local generation accuracy but also by document-level organization. This chapter formulates the main research questions and explains why discourse structure provides a promising avenue for addressing these constraints. This chapter also outlines the theoretical and empirical considerations that guide the investigation, thereby establishing the conceptual scaffolding for the chapters that follow.

Chapter 2 provides the technical background necessary for the dissertation. This chapter reviews foundational work on discourse structure, rhetorical relation modeling, and long-context language modeling, and surveys evaluation strategies commonly used in natural language generation, including metrics for coherence, salience, and factuality. This chapter clarifies the assumptions and limitations underlying the subsequent modeling choices, with particular attention to how discourse information is represented and learned.

Chapter 3 through **Chapter 7** progressively develop and evaluate a discourse-guided generation framework under increasingly complex conditions. **Chapter 3** explores Research Question 1 by introducing uncertainty-aware discourse representations for long-document summarization and examining how probabilistic rhetorical signals influence content selection and ordering. **Chapter 4** investigates Research Question 2 by extending discourse integration to parameter-efficient adaptation settings, analyzing whether structural guidance remains effective under constrained update capacity. **Chapter 5** expands the scope to multimodal scientific video-to-text summarization, corresponding to Research Question 3, and proposes a discourse-aware planning framework to support macro-level organization and factual grounding under fragmented evidence. **Chapter 6** addresses Research Question 4 by modeling explanatory content as a controllable discourse element in public-facing summarization, with a focus on balancing accessibility and factual reliability. **Chapter 7** examines Research Question 5 by extending discourse structure modeling to the retrieval-augmented generation setting, constructing intra-chunk RST trees and inter-chunk rhetorical graphs over retrieved evidence, and introducing a discourse-driven planning mechanism to guide answer generation across knowledge-intensive QA and summarization benchmarks.

Chapter 8 synthesizes the findings across the preceding chapters. This chapter revisits the five research questions and their associated hypotheses, identifying which aspects of discourse modeling generalize across settings and which remain sensitive to specific constraints. This chapter concludes by outlining open problems and directions for future research.

In addition to the main chapters, this dissertation includes references and appendices that provide the scholarly context and supplementary material supporting the arguments. The appendices contain implementation or experimental details that are closely related to the narrative but would disrupt the flow if included in the main text. This separation is intended to keep the primary chapters focused on the problems and evidence, while retaining the information necessary for careful inspection.

1.5 Published Work

The structure presented in the preceding section reflects the integration of several peer-reviewed publications developed during the course of the doctoral research. **Chapter 3** through **Chapter 7** correspond directly to five papers published in top-tier venues. These chapters preserve the core formulations, empirical findings, and methodological contributions of the original works, while the terminology, notation, and exposition have been harmonized to ensure a coherent dissertation-level narrative. The publications associated with **Chapter 3** through **Chapter 7** are listed below.

- **Chapter 3: 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), Toronto, Canada. Association for Computational Linguistics.
- **Chapter 4: 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), Mexico City, Mexico. Association for Computational Linguistics.

- **Chapter 5: Dongqi Liu**, Chenxi Whitehouse, Xi Yu, Louis Mahon, Rohit Saxena, Zheng Zhao, Yifu Qiu, Mirella Lapata, and Vera Demberg. 2025. *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), Vienna, Austria. Association for Computational Linguistics.
- **Chapter 6: Dongqi Liu**, Xi Yu, Vera Demberg, and Mirella Lapata. 2025. *Explanatory Summarization with Discourse-Driven Planning*. Transactions of the Association for Computational Linguistics, Volume 13.
- **Chapter 7: Dongqi Liu**, Hang Ding, Qiming Feng, Jian Li, Xurong Xie, Zhucun Xue, Chengjie Wang, Jiangning Zhang, Yabiao Wang. 2026. *Disco-RAG: Discourse-Aware Retrieval-Augmented Generation*. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), San Diego, USA. Association for Computational Linguistics.

Beyond the core technical chapters, two additional publications produced during the doctoral period contributed to the development of the overall research trajectory. While not included as standalone chapters, these works informed the problem formulation, dataset construction, evaluation methodology, and the conceptual framing of controllable generation that underpin this dissertation.

- **Dongqi Liu** and Vera Demberg. 2023. *ChatGPT vs Human-authored Text: Insights into Controllable Text Summarization and Sentence Style Transfer*. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), 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), Torino, Italy. ELRA and ICCL.

1.6 Statement of Individual Contributions

This dissertation is a cumulative dissertation. For all six publications listed above, the doctoral candidate (Dongqi Liu) led the entire research process, including the formulation of the research questions, the design of the methodologies, the implementation of all models, the execution and analysis of experiments, and the writing of the manuscripts. The candidate is also responsible for integrating the individual publications into a coherent dissertation narrative, including the writing of the introduction, background, and conclusion chapters, as well as the harmonization of terminology, notation, and exposition across all chapters.

Vera Demberg, as doctoral supervisor, provided continuous supervision, guidance, and feedback throughout all publications. Vera plays an instrumental role in shaping the overarching research direction, offering critical perspectives on experimental design and reviewing all manuscript drafts. Mirella Lapata co-supervised [Chapter 5](#) and [Chapter 6](#) during the candidate's research visit at the University of Edinburgh, contributing guidance on research methodology, experimental rigor, and manuscript presentation. Both supervisors contribute to refining the research questions and strengthening the argumentation across the respective publications.

All remaining co-authors provided discussions and supporting assistance at various stages of the respective projects, including but not limited to data collection, annotation, evaluation, and feedback on experimental design. Their contributions are acknowledged in the respective publications and in the acknowledgments section of this dissertation.

2

Background

This chapter establishes the foundational background needed for the remainder of this dissertation. We begin by formalizing the evidence-conditioned generation problem and introducing mainstream text generation paradigms, covering sequence-to-sequence architectures and decoder-only models, while highlighting long-input and attention constraints that are most relevant to the settings studied in this dissertation (Section 2.1). Next, we present two key discourse frameworks, Rhetorical Structure Theory (RST) and Question Under Discussion (QUD), as theoretical tools for modeling coherence and structure in text (Section 2.2). Building on this, we review parameter-efficient adaptation strategies and argue that discourse-level signals offer valuable guidance when model updates are constrained (Section 2.3). We also introduce retrieval-augmented generation as a complementary paradigm for knowledge-intensive scenarios (Section 2.4). We further discuss planning as an explicit mechanism for separating organization decisions from surface realization (Section 2.5). We finally conclude with a discussion of evaluation perspectives, focusing on output quality, discourse coherence, and factual consistency (Section 2.6).

2.1 Problem Setting and Task Scope

This section defines the problem setting addressed throughout this dissertation and clarifies the scope of what counts as evidence, what counts as acceptable outputs, and what kinds of failures matter in downstream applications. We first formalize the evidence-conditioned generation task and emphasize the gap between token-level plausibility and

discourse-level adequacy (Section 2.1.1). Next, we introduce two mainstream generation paradigms, examining how encoder–decoder architectures and decoder-only language models condition on the source input, and how training objectives and decoding strategies impact evidence accessibility (Section 2.1.2). We then discuss the challenges posed by long inputs, particularly in cases where models demonstrate strong local fluency but struggle with cross-paragraph coherence (Section 2.1.3).

2.1.1 Evidence-Conditioned Generation

Natural language generation (e.g., summarization) from complex evidence is commonly formulated as a conditional generation task, where a neural model parameterized by θ defines a distribution $p_\theta(y \mid x)$ over generated output y given an input x (Pilault et al. 2020; Zhong et al. 2025a). In standard summarization settings, x is usually a long-form text document, and the goal is to generate a concise and faithful summary that captures its central content. However, real-world applications often extend beyond this basic formulation. In particular, the input evidence may span multiple modalities (e.g., speech, images, or video) or the output may be expected to fulfill additional communicative goals, such as enhancing accessibility for non-expert audiences.

For example, in scientific video-to-text summarization, the input evidence can be expressed as a triplet $x = (x^{\text{text}}, x^{\text{audio}}, x^{\text{vision}})$, where text transcripts, audio narration, and slide visuals offer complementary but non-identical sources of support for the claims being summarized (Liu et al. 2025a; HangChen et al. 2025). In explanatory summarization, the target distribution may also incorporate explanatory additions that improve accessibility without altering the intended meaning of the original content (Joseph et al. 2024).

Across all settings, it is important to distinguish between sentence-level fluency and discourse-level adequacy. A generated answer may exhibit locally well-formed yet fail to convey meaningful content if, for instance, it detaches a result from the assumption that justifies it, flattens a contrastive structure to an undifferentiated enumeration, or omits a limitation that critically shapes the interpretation of a claim (Joseph et al. 2024). These failures are often masked at the surface level but become apparent when the generated answer is interpreted as an argument with internal dependencies.

These challenges are further amplified in the presence of long inputs, where the conditional distribution $p_\theta(y \mid x)$ is implemented through an interface with restricted

evidence access (Dong et al. 2024). In many previous studies, models operate on a transformed representation of x that reflects limitations imposed by context length, attention mechanisms, and preprocessing strategies. More formally, the models often condition on a reduced representation $\tilde{x} = \tau(x)$, where $\tau(\cdot)$ denotes operations such as truncation, segmentation, or retrieval, resulting in generation from $p_{\theta}(y | \tilde{x})$ rather than the ideal $p_{\theta}(y | x)$ (Luo et al. 2024). When $\tau(\cdot)$ excludes distant premises, background definitions, or earlier qualifications, the generated answer may drift toward statements that are superficially plausible but weakly grounded.

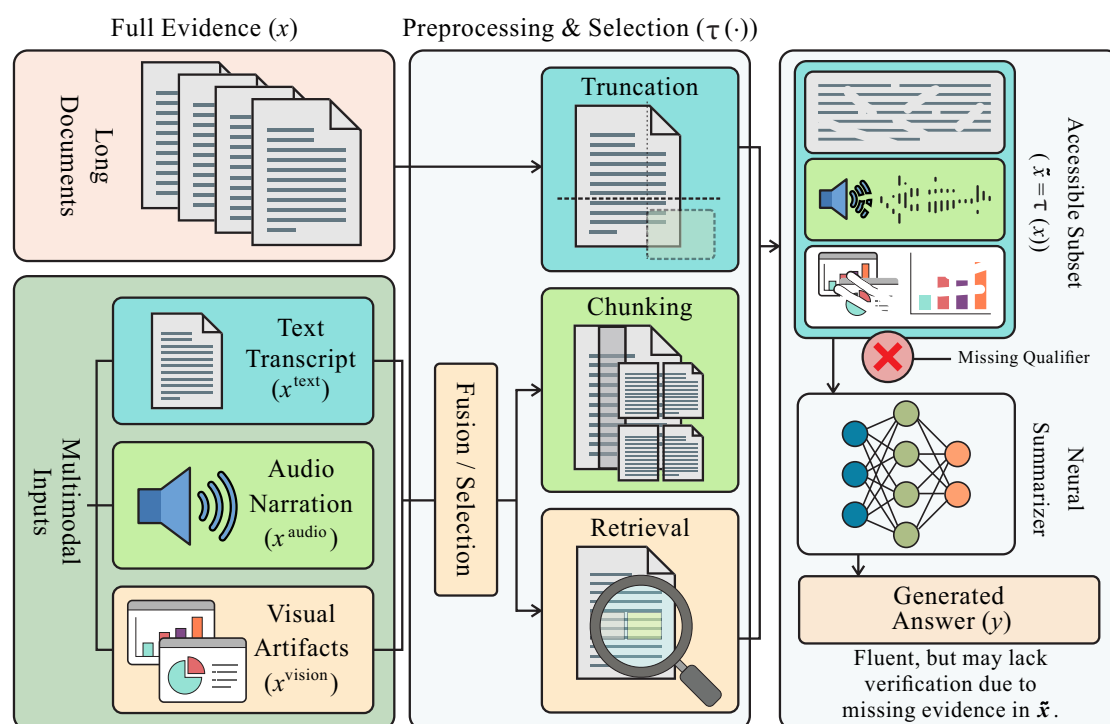


Figure 2.1: Evidence-conditioned generation under limited evidence access.

Figure 2.1 visualizes this formulation of conditional generation under constrained evidence access. By distinguishing the full input x from the accessible subset $\tilde{x} = \tau(x)$, Figure 2.1 underscores how preprocessing decisions can remove discourse-critical information before generation. In light of these limitations, this dissertation adopts a view of generation adequacy that emphasizes the grounding of claims in verifiable source content and the preservation of qualifiers, boundary conditions, and uncertainty markers wherever they are essential for interpretation (Wanner et al. 2025).

2.1.2 Model Paradigms for Text Generation

Figure 2.2 illustrates the two primary model paradigms that dominate current text generation research. The first paradigm consists of encoder-decoder sequence-to-sequence models (Lewis et al. 2020a; Zhang et al. 2020a), in which an encoder transforms the source input into contextual representations and a decoder generates the output while attending to the encoded output (see Figure 2.2 (a)). These kinds of models are typically trained using a maximum likelihood objective with teacher forcing, which can be written as:

$$\log p_{\theta}(y | x) = \sum_{t=1}^{|y|} \log p_{\theta}(y_t | y_{<t}, x). \quad (2.1)$$

This factorization underscores that conditioning on x is mediated through an internal representation learned by the encoder, along with cross-attention mechanisms in the decoder. The inductive bias of this architecture is that the source input is presented as an explicit memory, which can be queried at each decoding step. This characteristic makes encoder-decoder models a natural fit for the long-document settings explored in Chapter 3 and Chapter 4, both of which leverage explicit encoder-side computations and attention mechanisms that can be steered using discourse-level signals.

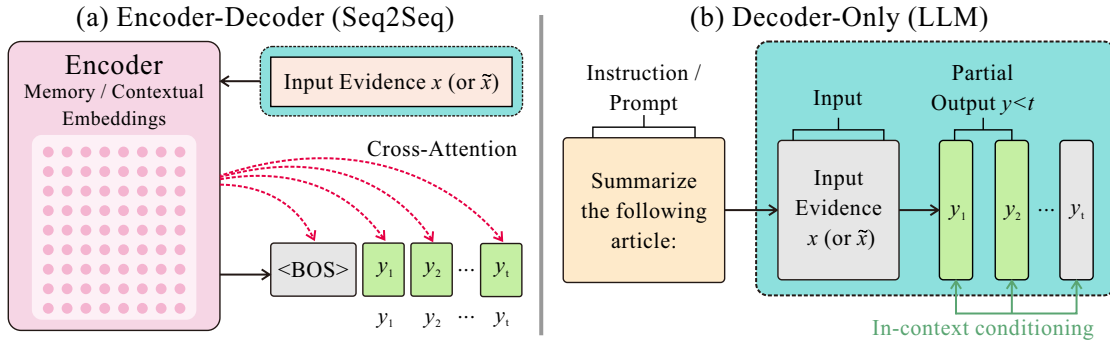


Figure 2.2: Comparison of two text generation paradigms.

The second paradigm comprises decoder-only large language models (see Figure 2.2 (b)), which treat text generation as a next-token prediction task conditioned on a context that includes both the source input and an instruction (Brown et al. 2020; Wei et al. 2022). In this paradigm, the boundary between source content and task instructions is encoded implicitly through input formatting, and the model learns to interpret different segments of the context. While the same conditional generation objective applies, conditioning

is realized through in-context representations rather than an explicit encoder–decoder interface. This paradigm is particularly relevant for instruction-following settings and multi-step prompting strategies, as well as for the planning-based pipelines introduced in [Chapter 5](#) and [Chapter 6](#), where the model first generates an explicit plan and then produces an output answer conditioned on that intermediate plan.

2.1.3 Long Inputs and Attention Patterns

Long scientific articles, multi-section reports, and multimodal evidence packages often distribute premises, caveats, and definitions at considerable distances from the claims they support ([Dao et al. 2022](#)). This poses a challenge for neural generation models, with the computational pressure arising from the attention mechanism. In standard Transformer architectures, full self-attention scales quadratically with sequence length (see [Figure 2.3](#) (a)), rendering direct application to long documents computationally prohibitive.

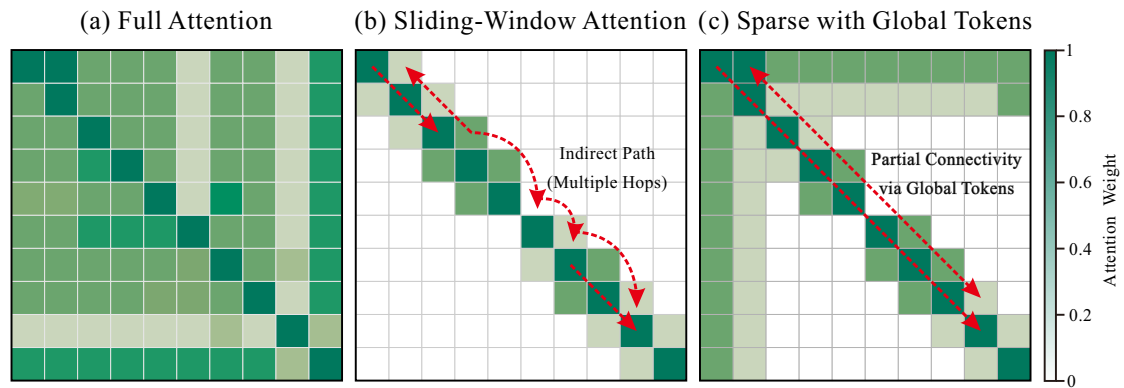


Figure 2.3: Comparison of different attention patterns.

For this reason, many long-context models adopt sparse attention patterns that restrict the set of positions each token can attend to. An often-used design employs local attention windows ([Zaheer et al. 2020](#)), in which each token attends only to a fixed-size neighborhood (see [Figure 2.3](#) (b)). While this design preserves strong local modeling capacity and improves scalability, it can weaken the model’s ability to capture long-range discourse dependencies across distant segments. For instance, when a premise is introduced early in the document, and a corresponding conclusion appears much later, a purely local attention mechanism may require a long sequence of intermediate attention

steps to propagate that dependency. This transmission path is vulnerable to disruption by topic shifts, section boundaries, or intervening boilerplate content.

A compromise involves introducing sliding-window attention augmented with a limited set of global tokens (see [Figure 2.3 \(c\)](#)). While global tokens offer a mechanism for long-range connectivity, their small number and tendency to aggregate information in a shallow and undifferentiated manner make it challenging to capture nuanced or implicit discourse connections ([Zaheer et al. 2020](#); [Tay et al. 2021](#); [Choromanski et al. 2021](#)). These properties help explain why models operating on long inputs often produce answers that are locally fluent yet fail to accurately reflect the document’s global structure and argumentative flow.

2.2 Discourse Structure

This section provides prerequisite knowledge on discourse structure. We first introduce Rhetorical Structure Theory as a framework for segmenting text into elementary discourse units (EDUs) and for labeling coherence relations along with their associated centrality roles ([Section 2.2.1](#)). We then present the Question under Discussion framework that conceptualizes discourse as a hierarchical sequence of implicit questions and answers, capturing the underlying information-seeking structure that governs discourse progression and coherence ([Section 2.2.2](#)).

2.2.1 Rhetorical Structure Theory

Rhetorical Structure Theory (RST) is a discourse framework for modeling the internal organization of a text. It describes how clauses and sentences are connected through rhetorical relations and structured into a coherent hierarchy ([Marcu. 1997](#); [Marcu. 1999b](#); [Marcu. 2000](#); [Kikuchi et al. 2014](#); [Goyal et al. 2016](#); [Liu et al. 2023c](#)). To be specific, RST delineates a set of coherence relations between textual segments, known as Elementary Discourse Units (EDUs), at the document level (e.g., one EDU might provide clarification for another, or conversely, two EDUs could present contrasting viewpoints). In addition to labeling the type of coherence relation, RST assigns centrality roles to EDUs based on their relative importance, distinguishing between *nuclei*, which convey the central content of the relation, and *satellites*, which provide supporting or supplementary information ([Marcu. 1999b](#); [Isonuma et al. 2019](#); [Xu et al. 2020](#)).

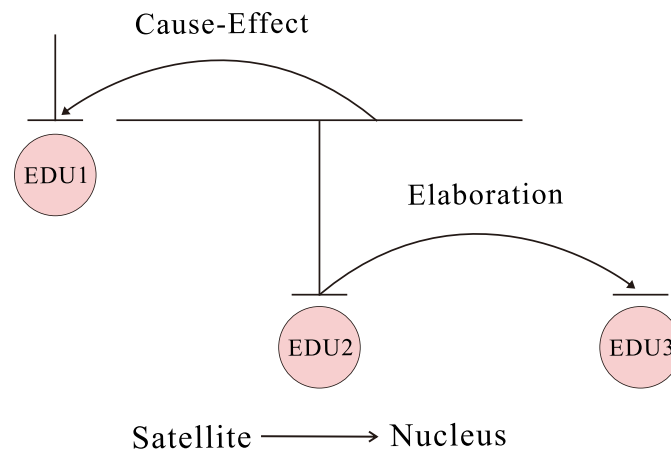


Figure 2.4: An example of RST tree: [Utilizing discourse structure to enhance text summarization is beneficial.]^{EDU1} [This technique can be used to identify key ideas and capture often overlooked nuances.]^{EDU2} [Accurate capture of these complex structures facilitates the generation of good summaries.]^{EDU3}

As an illustrative example, consider the RST tree depicted in [Figure 2.4](#). In this instance, EDU1 functions as the most pivotal component within the entire example, thus constituting the nucleus for both EDU2 and EDU3. EDU2 is tasked with elucidating and providing supplementary information to EDU3, positioning it as a satellite unit in relation to EDU3. Given the relatively lower discourse importance of EDU2, one could merge EDU1 and EDU3 while pruning EDU2; the semantic essence of the example would remain intact. In a more extreme scenario, retaining only EDU1 as the core sentence and omitting both EDU2 and EDU3, the primary information conveyed by the entire example would still be preserved. This reflects a broader observation in the literature ([Marcu. 1997](#); [Louis et al. 2010](#); [Cohan et al. 2018](#); [Liu et al. 2019a](#); [Li et al. 2020c](#); [Xu et al. 2020](#); [Dong et al. 2021](#); [Chen et al. 2021](#)) that *satellite* EDUs play a subordinate role, with the majority of the core sentences deriving from *nucleus* EDUs.

Automatic RST discourse parsing aims to recover the full discourse tree of a document and usually involves two main sub-tasks. The first is EDU segmentation, which identifies the minimal discourse units in a text. The second is tree construction, which determines the hierarchical structure, nuclearity assignments, and rhetorical relation labels among these units. Earlier neural approaches adopted transition-based or chart-based strategies and relied on pretrained encoder models such as BERT and XLNet to obtain span representations ([Kobayashi et al. 2020](#); [Liu et al. 2020](#); [Nguyen et al. 2021](#);

Koto et al. 2021b). More recent work introduced joint frameworks that integrate EDU segmentation directly into the tree parsing process, which eliminates the need for a separate preprocessing step and improves domain generality (Liu et al. 2021e). Yu et al. (2022) further proposed a second-stage EDU-level pre-training approach with two novel objectives, namely next EDU prediction and discourse marker prediction, that improve parsing accuracy on benchmark datasets. From 2023 onward, the field has witnessed several notable developments. Maekawa et al. (2024) demonstrated that decoder-only large language models can achieve state-of-the-art RST parsing performance by converting top-down and bottom-up parsing strategies into prompts and fine-tuning LLaMA with QLoRA. Their bottom-up variant achieves the best results on three benchmark datasets (RST-DT, Instr-DT, and GUM) and exhibits strong cross-corpus generalizability. On the theoretical side, Zeldes et al. (2025) introduced Enhanced RST (eRST), a framework that extends traditional RST to support discourse relation graphs with tree-breaking, non-projective, and concurrent relations, along with explicit signal annotations that provide interpretable rationales for discourse analyses. This extension is accompanied by a freely available English corpus of over 200K tokens across 12 spoken and written genres. Cross-lingual RST parsing has also advanced, with Chistova (2024) introducing a bilingual annotation and end-to-end parsing that demonstrates effective transfer even with limited second-language data. Despite these advances, RST parsers still exhibit reduced accuracy on out-of-domain data and long documents (Atwell et al. 2022; Liu et al. 2021e), which motivates the uncertainty-aware approach adopted in this dissertation.

2.2.2 Question Under Discussion

Question under Discussion (QUD) provides another perspective on discourse coherence by modeling discourse in terms of the implicit questions that are raised and resolved over time (Beaver et al. 2017; Wu et al. 2023a). From this view, a sentence is perceived as relevant not merely because of topical similarity, but because it addresses an information need that has become salient in the evolving discourse context. Discourse progression can therefore be understood as a structured process in which a question is introduced or made prominent, followed by an answer that updates the common ground and shifts what is currently under discussion (Wu et al. 2023a).

Figure 2.5 illustrates this idea with an example. Given the statements about Antarc-

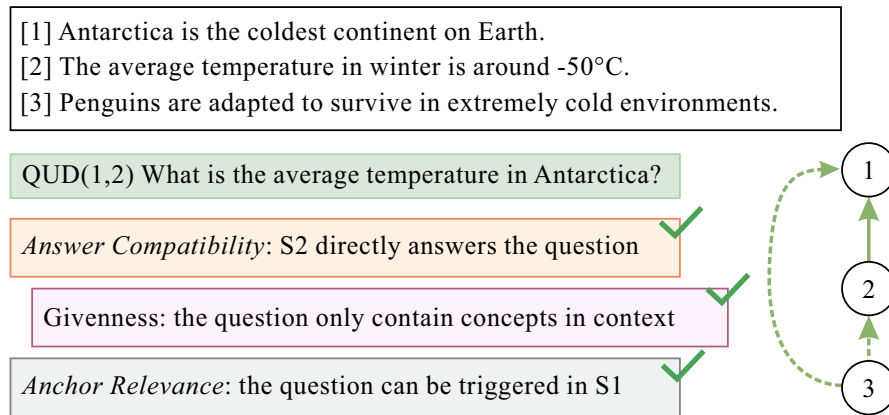


Figure 2.5: An illustration of the Question under Discussion (QUD) framework.

tica, the initial sentence establishes a salient topic that licenses the implicit question “What is the average temperature in Antarctica?”. The second sentence directly answers this question, satisfying answer compatibility, givenness, and anchor relevance, while the third sentence, although topically related, does not directly resolve the active question. This example highlights how QUD captures relevance in terms of question–answer alignment rather than surface proximity, and how it distinguishes between sentences that directly advance the discourse and those that introduce parallel or subsequent informational threads. This perspective is particularly useful for text generation tasks, as it shifts attention from lexical overlap or sentence-level importance to communicative function. From a QUD standpoint, a high-quality summary can be viewed as a sequence of answers to the most important questions posed by the source document, with sentence ordering guided by a coherent progression of those questions (Pagnoni et al. 2023; Wu et al. 2023b).

While QUD has traditionally been studied through manual annotation, recent years have seen growing efforts to automate the derivation of QUD structures. De Kuthy et al. (2020) proposed an early approach that combines transformation rules with a neural question generation model to produce candidate QUD questions for a given sentence. Ko et al. (2023) developed the first QUD dependency parser capable of deriving full-document question structures, trained on the large crowdsourced DCQA dataset. Wu et al. (2023a) introduced QUDeval, the first systematic evaluation framework for QUD parsing, and showed through fine-grained analysis of 2,190 generated questions that satisfying all theoretical constraints of QUD remains challenging for both fine-tuned systems and large language models. On this basis, Suvarna et al. (2024) proposed QUD-

SELECT, a joint-training framework that uses instruction-tuning and selective decoding to simultaneously predict anchor sentences and generate associated questions. This approach achieves improvements over prior pipelined methods in both human and automatic evaluations. More recently, Han et al. (2025a) adopted reinforcement learning with the Group Relative Policy Optimisation (GRPO) objective for LLM post-training on QUD generation and achieves higher-quality outputs than supervised fine-tuning baselines with only three annotated exemplars. On the annotation and modeling front, Fu et al. (2025) proposed a topicality-driven QUD model that achieves higher inter-annotator agreement than prior QUD tree approaches and compares its annotations with RST structures. Fu (2025) provided a comprehensive survey of QUD models for discourse processing that examines the relationship between QUD and mainstream discourse frameworks such as RST, PDTB, and SDRT. Despite these advances, the automatic derivation of QUD structures that fully satisfy all theoretical constraints remains an open challenge. This dissertation leverages QUD not as a target for full structural parsing but as a planning mechanism that guides text generation through explicit intermediate question sequences.

2.3 Resource Constraints and Efficient Adaptation

This section provides background on efficient adaptation strategies for neural generation models, with an emphasis on the constraints that arise when inputs are long and model backbones are large. We first present three common parameter-efficient fine-tuning (PEFT) approaches and discuss why memory and computational resources often become the bottleneck factors when adapting models to new domains (Section 2.3.1). We then discuss how such constraints on update capacity complicate the learning of discourse-level behaviors (Section 2.3.2).

2.3.1 PEFT Methods

Parameter-efficient fine-tuning (PEFT) is motivated by pressure in current model development. While large pretrained language models provide strong initial representations, achieving further performance gains through full fine-tuning is often computationally prohibitive (Hu et al. 2022; Li et al. 2021). This difficulty is amplified in long-context generation, where training requires storing large activation tensors for backpropagation

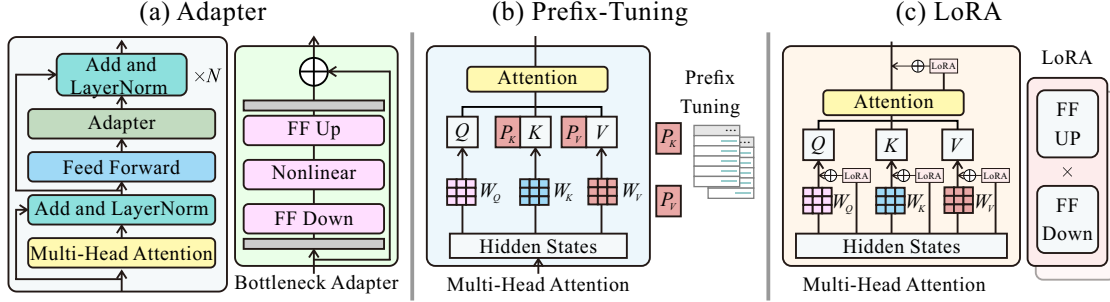


Figure 2.6: Overview of parameter-efficient adaptation.

and maintaining optimizer states for all trainable parameters. In such settings, the dominant constraint is GPU memory rather than wall-clock time, which directly limits the feasible batch sizes and maximum sequence lengths that can be processed (Hu et al. 2022). These constraints have led to the development of PEFT strategies that keep the majority of the model backbone frozen while introducing a lightweight set of trainable parameters that are more memory-efficient and easier to iterate on.

PEFT methods can be categorized according to which components of the computation they modify. As shown in Figure 2.6, adapter-based approaches insert small feed-forward modules between Transformer layers and train only these additional modules while keeping the original projection weights frozen (Rücklé et al. 2021; Karimi Mahabadi et al. 2021). Prompt- and prefix-based methods learn additional vectors that are injected as extra conditioning signals at the input or within attention (Lester et al. 2021; Li et al. 2021). Among these approaches, Low-rank adaptation (LoRA) has been widely adopted due to its simplicity and explicit control over update capacity while remaining compatible with standard linear projections. In LoRA, an existing weight matrix $W \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ is kept fixed, and adaptation is achieved by adding a learned update ΔW , such that the effective weight becomes $W + \Delta W$. The update ΔW is parameterized as a product of two low-rank matrices (Hu et al. 2022). A typical parameterization writes:

$$\Delta W = BA, \quad B \in \mathbb{R}^{d_{\text{in}} \times r}, \quad A \in \mathbb{R}^{r \times d_{\text{out}}}, \quad r \ll \min(d_{\text{in}}, d_{\text{out}}). \quad (2.2)$$

Quantization-aware variants further reduce memory consumption by storing the frozen backbone weights in low-precision formats while training a small set of adaptation parameters in higher precision (Dettmers et al. 2023). This design significantly lowers the memory footprint of large models, making fine-tuning feasible in scenarios where storing the full-precision backbone would exceed hardware constraints (see Figure 2.7).

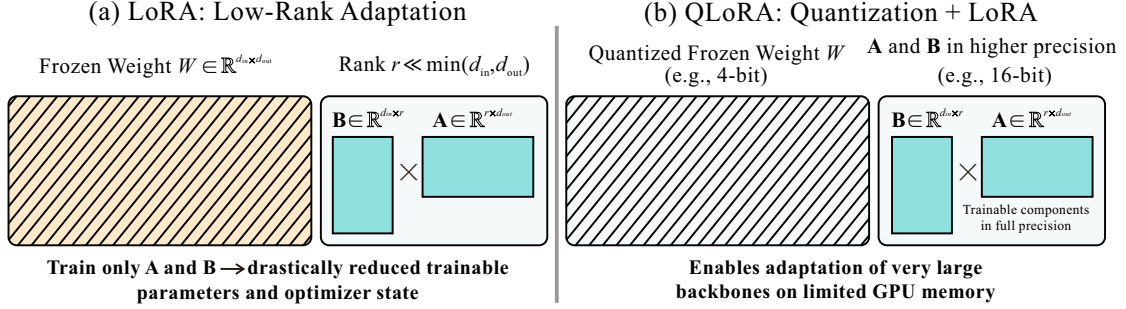


Figure 2.7: Illustration of low-rank adaptation and quantization-aware variants.

2.3.2 Challenges of PEFT Methods

Long-form generation exacerbates resource constraints in ways that go beyond the sheer number of trainable parameters. A restricted update budget often coincides with a weak training signal for learning global organization (Phang et al. 2023). When update capacity is small, performance gains tend to manifest primarily in surface-level realization, while document-level salience and ordering patterns remain largely unaffected (Ravaut et al. 2024; Jung et al. 2025). This limitation is consequential for generation tasks that require discourse-level organization, where many failure cases are not from grammatical errors but stem from structural misinterpretation (Liu et al. 2024a). These observations motivate the incorporation of structural cues during adaptation, so that limited parameter updates are preferentially allocated toward evidence selection and discourse organization.

Discourse cues are usually derived from automatic parsers and thus should be treated as uncertain. RST parsers segment documents into EDUs and assign relation labels that characterize the rhetorical connections between spans. Rather than treating parser output as a single discrete tree, a more robust strategy is to preserve uncertainty and represent discourse information as a soft distribution (Liu et al. 2023c; Liu et al. 2024a). This uncertainty-aware choice aligns well with PEFT settings. When update capacity is limited, imposing incorrect hard constraints can be difficult to override through learning. A distributional representation allows the model to exploit discourse signals when they are informative, while allowing the generation objective to discount ambiguous or misleading cues.

2.4 Retrieval-Augmented Generation

This section introduces retrieval-augmented generation (RAG) for knowledge-intensive text generation, where the information needed to produce a faithful response may not be fully contained within the model’s parametric memory or the immediately available context. We first describe the standard retrieve-then-generate framework, covering retrieval mechanisms, dense encoding strategies, and recent architectural extensions such as iterative and adaptive retrieval (Section 2.4.1). We then discuss the structural limitations that arise when retrieved evidence is treated as a flat concatenation of independent chunks, and review emerging efforts to incorporate discourse-level organization into the retrieval and generation process (Section 2.4.2).

2.4.1 The Retrieve-then-Generate Paradigm

In many knowledge-intensive scenarios, the information needed to produce a faithful response is not fully contained within the provided context or the model’s parametric memory. Retrieval-augmented generation (RAG) addresses this gap by introducing an explicit retrieval step that gathers relevant passages from an external corpus before generation (Lewis et al. 2020b; Guu et al. 2020). The core motivation is that parametric models, regardless of their scale, cannot memorize the entirety of world knowledge, and their internal representations may become outdated as new information emerges. By externalizing knowledge into a retrievable corpus, RAG systems decouple factual storage from language generation, allowing the model to access up-to-date and domain-specific evidence at inference time without retraining (Guu et al. 2020).

As illustrated in Figure 2.8, a standard RAG pipeline operates in several stages. The external corpus \mathcal{D} is first segmented into chunks, which are then encoded into dense vector representations and stored in an index. At query time, the system retrieves a set of chunks $\mathcal{C}(q; \mathcal{D}) = \{c_1, c_2, \dots, c_k\}$ whose embeddings are most similar to the query embedding, and the generator produces the output y by conditioning on both the query and the retrieved context, $p_\theta(y | q, \mathcal{C})$ (Lewis et al. 2020b; Karpukhin et al. 2020). The retrieval component itself can be realized in different ways. Early approaches rely on sparse lexical matching methods such as BM25 (Robertson et al. 2009), which score documents based on term-frequency statistics and are effective when the query and relevant passages share surface-level vocabulary. However, sparse retrieval struggles with vo-

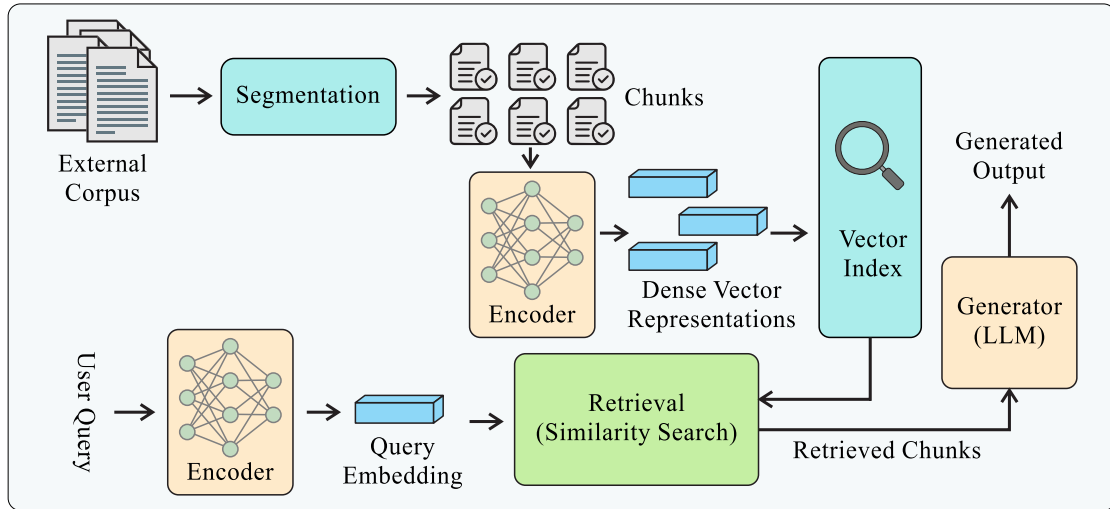


Figure 2.8: Standard retrieval-augmented generation pipeline. An external corpus is segmented into chunks, encoded into dense vectors, and indexed. At query time, the top- k chunks are retrieved based on embedding similarity and concatenated as context for the language model.

cabulary mismatch, where semantically relevant passages may use different terminology than the query. Dense retrieval models address this limitation by encoding both queries and passages into continuous vector spaces, where semantic similarity is measured via inner product or cosine distance (Karpukhin et al. 2020). Hybrid strategies that combine sparse and dense signals have also been explored and often yield complementary benefits (Ma et al. 2021).

This retrieve-then-generate paradigm has been shown to improve performance on knowledge-intensive tasks such as open-domain question answering, fact verification, and knowledge-grounded dialogue (Lewis et al. 2020b; Asai et al. 2024; Chan et al. 2024). Beyond single-round retrieval, several architectural extensions have been proposed to improve the coverage and relevance of retrieved evidence. Iterative retrieval refines the query based on intermediate generation steps, allowing the model to progressively gather more targeted evidence as its understanding of the task evolves (Jiang et al. 2023b). Adaptive retrieval introduces a decision mechanism that determines whether retrieval is needed at each generation step, avoiding unnecessary retrieval when the model’s parametric knowledge is sufficient to produce a confident prediction (Asai et al. 2024). Self-reflective RAG further extends this idea by training the model to critique its own retrieval decisions and generation outputs through learned reflection tokens that indicate whether retrieved passages are relevant, whether the generated response is sup-

ported, and whether the output is useful (Asai et al. 2024). Structure-aware variants have also been proposed, including graph-based methods that organize evidence into knowledge graphs to preserve entity-level and relational information (Edge et al. 2024; Sanmartin. 2024), and hierarchical approaches that dynamically select the most appropriate structural format for the retrieved evidence based on the nature of the query (Li et al. 2025c).

2.4.2 Challenges of RAG Methods

Despite these advances, standard RAG pipelines treat retrieved chunks as independent units that are concatenated in a flat sequence without modeling rhetorical connections across them. This design implicitly assumes that the generator can infer the appropriate relationships among chunks from their content alone, an assumption that frequently breaks down in practice.

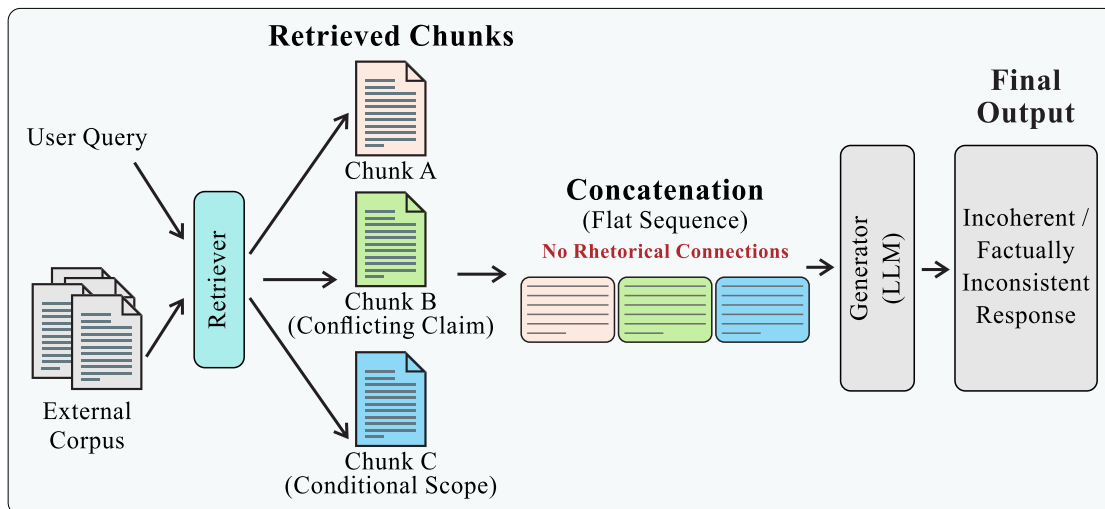


Figure 2.9: Structural limitations of standard RAG pipelines. Retrieved chunks are concatenated without explicit rhetorical connections. When chunks contain conflicting or conditionally scoped claims, the generator lacks the structural signals needed to reconcile them, which can lead to overgeneralization or factual inconsistency.

As illustrated in Figure 2.9, the flat representation of retrieved evidence introduces several failure modes. When two chunks present claims that hold under different conditions or scopes, the generator may merge them into a single overgeneralized statement that is not supported by either source individually. Similarly, when chunks offer contrasting perspectives, the absence of explicit contrastive or concessive relations

can cause the model to favor one perspective while ignoring the other, or to produce internally contradictory outputs. These issues are especially pronounced in multi-hop reasoning scenarios, where the generator should synthesize information across multiple passages to arrive at a coherent answer (Yang et al. 2018; Trivedi et al. 2023). Without structural signals that indicate how pieces of evidence relate to one another, the model may fail to chain supporting premises correctly or may introduce spurious inferential links.

Several recent efforts have begun to address this structural gap. Knowledge graph-augmented approaches construct structured representations over retrieved passages, linking entities and relations to facilitate multi-hop reasoning (Edge et al. 2024; Sanmartin. 2024). Chain-of-thought prompting strategies encourage the model to externalize its reasoning steps, which can partially compensate for the missing inter-chunk structure (Trivedi et al. 2023). However, these methods primarily operate at the entity or propositional level and do not capture the rhetorical functions that govern how claims relate to one another at the discourse level, such as whether one passage elaborates on, contradicts, or provides conditions for another.

2.5 Planning and Structured Generation

This section introduces planning as a mechanism for structured text generation. We adopt the view that a good output is not only a sequence of fluent sentences but also an ordered presentation of claims whose organization shapes how the content is interpreted. Such an organization is more readily controlled when an explicit plan is available. We first describe discourse-aware plans as intermediate targets that make macro-level structure explicit and provide an interface between decisions about what information to convey and how to realize it (Section 2.5.1). We then explain why planning becomes particularly valuable in settings where evidence is fragmented across modalities or when summaries incorporate explanatory content that should remain anchored to specific source claims (Section 2.5.2).

2.5.1 Plans as Explicit Intermediate Targets

Plan-based generation starts from an observation about long-form text generation. Many failures that matter to readers are not local language failures, but global organizational

failures (Liu et al. 2022d). These failures arise when the model foregrounds irrelevant or secondary claims, presents content in an implausible or incoherent order, or omits qualifiers that are necessary for interpreting an argument. Introducing an explicit plan provides a mechanism for making such global decisions before surface realization (Adams et al. 2023).

Formally, planning introduces an intermediate variable b that encodes macro-level structure. Given input evidence x , the model generates an output y . A plan-based model factorizes conditional generation by allowing the plan to mediate between evidence and output:

$$p_{\theta}(y | x) = \sum_b p_{\theta}(b | x) p_{\theta}(y | x, b). \quad (2.3)$$

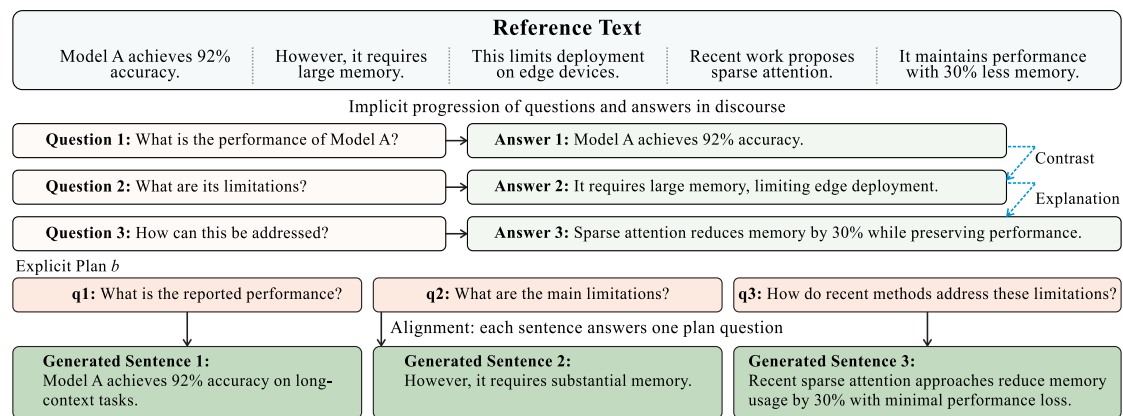


Figure 2.10: Planning-based generation in which a plan b is generated from evidence and then used to condition the final output.

As shown in Figure 2.10, the plan makes explicit the progression from evidence to an ordered sequence of output sentences. In the settings studied in this dissertation, a particularly effective instantiation represents the plan as a sequence of questions $b = (q_1, \dots, q_T)$, where each question q_t is answerable from the evidence x , and each output sentence corresponds to an answer. This representation aligns with the Question under Discussion (QUD) perspective introduced earlier. A well-formed summary can be viewed as a concise sequence of answers to the most salient information needs raised by the source (Wu et al. 2024b).

Beyond guiding organization, explicit plans also support controllability and post-hoc verification (see Figure 2.11). By editing the plan b , for example, by simplifying

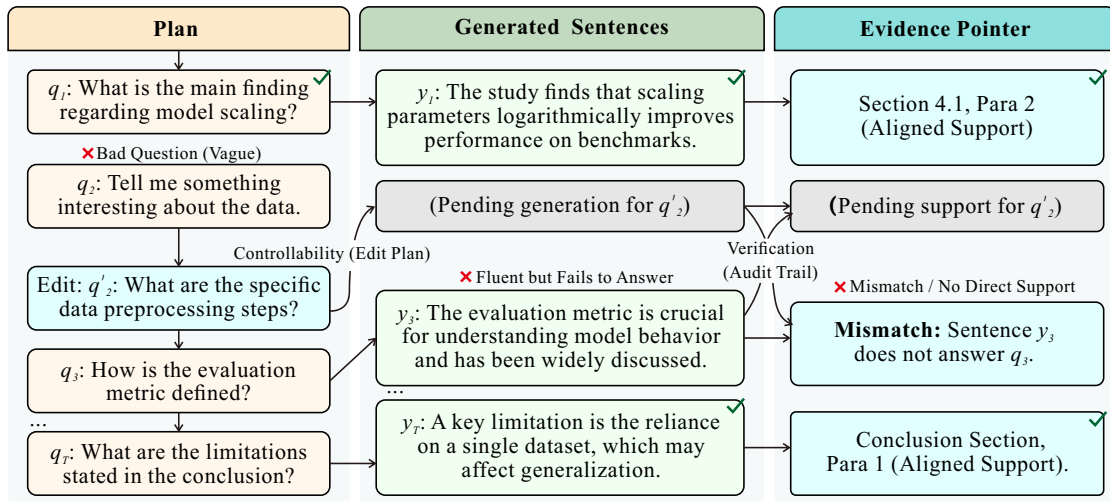


Figure 2.11: Planning as an interface for controllability and verification.

questions for non-expert readers or removing questions to increase compression, one can modify the resulting output in a targeted manner. Moreover, because each sentence is grounded in a corresponding question, the plan provides a transparent checklist of intended claims, thereby facilitating auditing and interpretation (Scialom et al. 2021).

2.5.2 Planning across Modalities and Audiences

Planning becomes particularly valuable when the evidence supporting an answer is distributed across multiple modalities. As illustrated in Figure 2.12, multimodal inputs may encode complementary forms of information. Core premises may be conveyed through speech, quantitative details through visual elements such as charts, and definitions through accompanying text or slides (Liu et al. 2025a). In such settings, models that generate directly from a fused multimodal representation can appear coherent while omitting a crucial qualifier that is present only in one modality. These errors can be especially difficult to detect, as the generated sentences remain contextually fluent and contextually plausible. Planning mitigates this fragility by explicitly making the informational pathway. Rather than relying on implicit decoding dynamics to infer modality-specific support for each detail, a plan articulates the intended function of each sentence and ensures that it is answerable from the available evidence.

Differences in the audience also shape what constitutes an appropriate plan. In expert-oriented summaries, plans can presuppose domain-specific knowledge and fo-

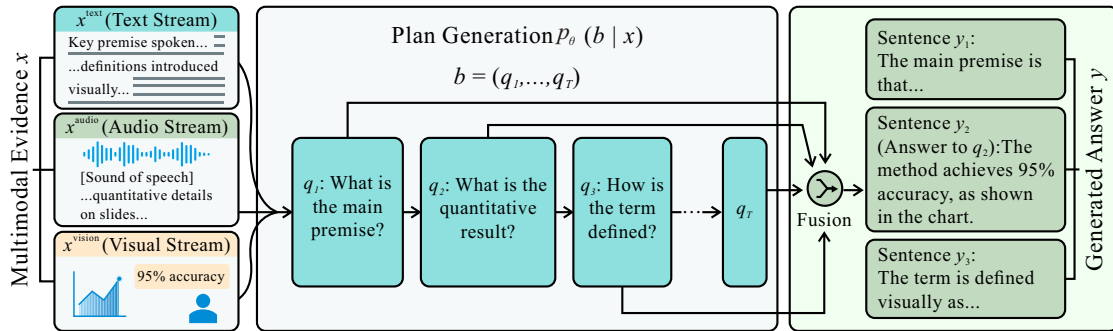


Figure 2.12: Planning as a mechanism for linking multimodal evidence to sentence-level intents.

cus on methodological rigor, empirical results, and nuanced limitations that are salient to specialists. In contrast, lay-oriented summaries require plans that foreground definitions, background context, and explanatory elements that support comprehension without altering the underlying claims (Wu et al. 2023b). A question-based planning framework renders this linkage concrete by treating explanatory sentences as answers to questions raised by preceding claims (Cardenas et al. 2023). It also enables control over where explanations appear, since the plan can specify which claims should trigger additional questions that invite clarification.

2.6 Evaluation and Verification

This section provides background on the evaluation of generation systems across the settings examined in this dissertation. Evaluation is inherently challenging because multiple outputs may be acceptable under different compression levels, audience assumptions, or task requirements, and sentence-level fluency alone does not guarantee that an output is informative, well-organized, or faithful to the underlying evidence. In this section, we first review commonly used automatic quality metrics (Section 2.6.1), and then discuss verification-oriented metrics that assess grounding and factual consistency, particularly in settings involving long inputs, multimodal evidence, or retrieved passages (Section 2.6.2). We further describe the human evaluation protocols adopted across all experimental chapters (Section 2.6.3), and introduce the LLM-as-a-judge paradigm as a scalable complement to human annotation (Section 2.6.4).

2.6.1 Automatic Quality Metrics

Automatic evaluation often begins with reference-based metrics that compare a generated answer y against one or more human-written references r . These metrics are widely adopted because they are relatively inexpensive to compute, reproducible across studies, and straightforward to aggregate at scale. They operationalize informativeness as similarity to a trusted target realization, while acknowledging that overlap-based comparisons are only an approximation because multiple phrasings can express the same content, and reference answers themselves may differ in content selection, level of detail, or ordering. With these limitations in mind, we review the reference-based metrics employed throughout this dissertation below.

Similarity and Coverage Metrics. Lexical and semantic similarity metrics aim to assess how well a generated answer y aligns with one or more human-written references r . This dissertation adopts several widely used reference-based metrics, each capturing a different facet of similarity or coverage under various assumptions of equivalence.

- **Exact Match (EM)** measures the proportion 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.
- **ROUGE (Lin. 2004)** calculates overlap between tokens derived from y and r . We report F1 variants of Rouge-1 (R1), Rouge-2 (R2), Rouge-L (RL), and Rouge-Lsum (RLsum) depending on the setting. Rouge-1 and Rouge-2 emphasize unigram and bigram overlap, while Rouge-L and Rouge-Lsum use longest common subsequence matching.
- **BERTScore (Zhang et al. 2020b)** measures semantic similarity by aligning contextual embeddings between y and r .
- **METEOR (Banerjee et al. 2005)** performs unigram-level matching with support for stemming, synonymy, and paraphrase mappings, and computes a recall-biased harmonic mean of precision and recall.
- **sacreBLEU (Post. 2018)** reports standardized BLEU-style n-gram precision using consistent tokenization and smoothing protocols, allowing reproducible comparisons across systems.
- **NIST (Lin et al. 2003)** extends BLEU by weighting n-grams based on their information content, assigning higher scores to rare, content-bearing expressions.

- CIDEr-D (Vedantam et al. 2015) computes TF-IDF-weighted n-gram similarity to reference answers, with length-based penalties that downweight repetitive or overly verbose matches.

Reference-based metrics provide useful signals for assessing coverage and phrasing similarity, but they are not designed to capture whether ordering decisions preserve document-level organization or whether interpretive qualifiers are retained. For this reason, this dissertation treats such scores as complementary indicators of informativeness with respect to the reference answers, rather than as standalone proxies for discourse coherence or answer quality.

Readability Metrics. When summaries are intended for non-expert audiences, it is important to assess readability and the extent to which explanatory content is included. This dissertation employs D-SARI (Sun et al. 2021) and the Flesch Reading Ease metric (FRE Kincaid et al. 1975) as readability-oriented measures, both reported on a 0–100 scale, with higher values indicating more accessible text.

- D-SARI extends the sentence-level SARI metric to the document level and evaluates the quality of simplification operations by measuring how well a system performs lexical additions, deletions, and retentions relative to both the source document and reference simplifications. It is particularly suited for document-level summarization tasks where simplification spans multiple sentences and structural reorganization may occur.
- FRE estimates readability based on two surface-level properties of the text, namely the average number of words per sentence and the average number of syllables per word. Texts with shorter sentences and fewer multi-syllabic words receive higher FRE scores, which indicates greater ease of reading.
- Chapter 6 introduces `ExpRatio`, defined as the proportion of explanatory EDUs among all EDUs in a generated summary. This metric serves as an indicator of how frequently the model inserts explanatory units in relation to central claim-bearing content.

2.6.2 Grounding and Factuality Metrics

Grounding and factuality assessments aim to evaluate whether the claims expressed in a generated answer are adequately supported by the available evidence. This dissertation

employs several grounding-oriented metrics selected to match the evidence modality of the source input and the predominant factual risks specific to each setting.

- SummaC (Laban et al. 2022) detects semantic inconsistency by decomposing source inputs and generated summaries into sentence-level units and aggregating entailment-style scores over all possible sentence pairs.
- VideoScore (He et al. 2024) provides fine-grained, multi-aspect alignment scores for evaluating consistency between textual answers and multimodal evidence composed of video, speech, and visual frames.
- FactVC (Liu et al. 2023d) quantifies factual consistency between generated text and video content by combining coarse-grained similarity and fine-grained matching, with values reported on a 0 to 100 scale.
- VeriScore (Song et al. 2024b) supports external verification by checking whether extracted claims from the generated answer can be corroborated using external knowledge sources.

While these automatic metrics provide useful quantitative signals, each category has notable limitations. Reference-based metrics such as ROUGE and BERTScore primarily capture lexical and semantic overlap with reference answers but cannot assess whether content is correctly ordered, whether cross-sentence transitions are coherent, or whether interpretive qualifiers are preserved. Readability metrics such as D-SARI and FRE operate on surface-level textual properties and do not evaluate whether explanations are accurate, well-placed, or serve appropriate discourse functions. Similarly, grounding and factuality metrics such as SummaC and VeriScore operate at the claim or sentence level and may miss document-level consistency issues, such as contradictions between distant paragraphs or factual distortions introduced by incorrect cross-sentence inference. More broadly, coherence failures often manifest as distortions in cross-sentence structure or logical flow that can persist even when local phrasing appears fluent and relevant. No single automatic metric is designed to capture all of these dimensions simultaneously.

2.6.3 Human Evaluation

To address the limitations of automatic metrics, this dissertation incorporates human evaluation across all experimental chapters. Human judges rate generated outputs across multiple quality dimensions tailored to each task setting. Common dimensions include

faithfulness, which assesses whether the generated answer accurately reflects the source content without introducing unsupported claims; *relevance*, which measures whether the selected content addresses the central topics of the source; *informativeness*, which evaluates the coverage of important details; *conciseness*, which captures whether the generated answer avoids unnecessary redundancy; and *coherence*, which assesses whether sentences follow a logical order and form a well-structured whole. For lay summarization (Chapter 6), two additional dimensions are introduced, namely *explanation accuracy* and *explanation usefulness*, which evaluate whether explanatory content is factually correct and whether it genuinely aids comprehension. Across all chapters, each generated answer is independently rated by annotators. However, human evaluation is inherently constrained by the time and cost of annotation. Due to the considerable length of source inputs in text generation from long-document and video settings, it is impractical to scale human evaluation through crowdsourcing.

2.6.4 LLM-as-a-Judge

To facilitate larger-scale evaluation while approximating human judgment, this dissertation also adopts the LLM-as-a-judge paradigm (Liu et al. 2023f; Zheng et al. 2023; Liusie et al. 2024). In this approach, a large (multimodal) language model is prompted to evaluate generated answers along the same dimensions used in human evaluation. Across all experimental chapters, LLM-based evaluators generally exhibit substantial agreement with human judges on relative ranking and score distributions. Nevertheless, several known limitations apply. LLM evaluators may exhibit biases inherited from pretraining, such as favoring certain surface-level stylistic patterns or penalizing unconventional but valid phrasings. Prior work has also shown that pairwise comparative assessment tends to outperform absolute scoring in terms of alignment with human preferences (Liusie et al. 2024; Liu et al. 2024d). These findings suggest that LLM-as-a-judge evaluation is a useful complement to human annotation but should not be treated as a standalone substitute, particularly for aspects that require deeper contextual reasoning.

Last but not least, this dissertation interprets automatic indicators, human evaluations, and LLM-based assessments as complementary perspectives on output quality. Automatic metrics provide efficient, reproducible coverage of surface-level properties. Human evaluations offer fine-grained judgments on dimensions that require contextual understanding, such as coherence, faithfulness, and explanation usefulness. LLM-as-a-

judge evaluations bridge the gap between the two by enabling larger-scale assessments that approximate human preferences, with the caveat that their judgments require validation against human ratings in each task setting.

2.7 Chapter Summary

This chapter has outlined the conceptual and technical foundations for studying text generation as an evidence-conditioned task under long-context and retrieval-augmented constraints. This chapter introduced discourse structure, parameter-efficient adaptation, planning mechanisms, and retrieval-augmented generation as components for reasoning about content selection, macro-level organization, explanation design, and evidence integration from external sources. This chapter also surveyed evaluation and verification strategies, emphasizing the importance of aligning quality metrics with grounding checks in settings involving extended textual inputs, multimodal sources, or independently retrieved passages. These perspectives provide the theoretical scaffolding for the models, methods, and empirical analyses presented in the subsequent chapters of this dissertation.

3

Discourse-Aware Attention for Text Summarization

The challenge of long-document summarization lies not only in length compression but, more crucially, in identifying which information is worth preserving across large spans of text. [Chapter 2](#) frames summarization as a conditional generation task under evidence-limited conditions and highlights the asynchrony between local readability and global document usability. In other words, when premises, constraints, and conclusions are distributed across distant positions in a long document, text generation models often access only partial information due to constrained input interfaces, which increases the likelihood of content selection errors and incoherent sentence transitions. Furthermore, while sparse attention mechanisms (e.g., sliding-window attention) enhance computational feasibility for long inputs, they may also weaken the explicit pathways that transmit inter-paragraph dependencies. Against this backdrop, Research Question 1 (RQ1) focuses on how to transform discourse structure into an uncertainty-aware signal that can directly inform content selection and ordering decisions, instead of one that is relegated to post-processing or constrained by a static parse tree.

To this end, [Chapter 3](#) tests Hypothesis I that a distributional representation of discourse structures, rather than a single-best parse, may help mitigate information loss caused by erroneous or overconfident discourse labels. This, in turn, enables the summarization model to make more robust decisions about content organization under structural ambiguity and parsing noise. Concretely, [Chapter 3](#) adopts Rhetorical Structure Theory (RST) to represent discourse relations and proposes to transform n-best dis-

course parses, along with their confidence scores, into a Labeled Discourse Distribution (LDD) tensor. This tensor is then aligned and injected into the Longformer model’s sliding-window attention computation through soft modulation of the attention scores. This design introduces discourse-aware inductive bias and preserves the computational structure of long-input encoding. Subsequent experiments on long-document summarization datasets such as BookSum, eLife, and Multi-LexSum evaluate the utility of this distributed discourse signal across dimensions such as content coverage, coherence, and faculty consistency. [Chapter 3](#) also compares variants that incorporate only structural uncertainty with those that include relation-type distributions and analyzes their relative impact on model performance and behavior.

3.1 Problem Formulation

For writing a good summary of a long document, it is of paramount importance to discern the salient information within the text and to comprehend the intricate interconnections among its various components. Contemporary leading-edge systems for abstractive (long) text summarization employ Transformer ([Vaswani et al. 2017](#)) encoder-decoder architecture ([Zaheer et al. 2020](#); [Guo et al. 2022](#)). These sequence-to-sequence models first transform the source document into a high-dimensional content representation and then decode the predicted summary conditioned on the representation ([Belinkov et al. 2018](#); [Xu et al. 2019](#); [Cao et al. 2022b](#); [Balachandran et al. 2021](#)). It has been demonstrated in the past that such an architecture does a poor job of digging high-level discourse structure during the encoding phase ([Lin et al. 2019](#); [Zhang et al. 2020c](#); [Koto et al. 2021a](#); [Wynter et al. 2023](#)). However, discourse structure is important for deciding what to include vs. not to include in the summary ([Marcu. 1997](#); [Marcu. 1999b](#); [Marcu. 1998](#); [Zhong et al. 2020](#)). Given that previous work has indicated that the performance of neural language models can be enhanced through the incorporation of latent structure information ([Ettinger. 2020](#); [Qian et al. 2021](#); [Liu et al. 2022b](#)), we will here explore the integration of discourse relation structure into the Longformer model ([Beltagy et al. 2020](#)); this architecture has been shown to be particularly suitable for encoding long input texts.

Rhetorical Structure Theory (RST) serves as a discourse framework designed to articulate the interrelationships among sentences at the document level. This framework distinguishes a plethora of coherence relations delineating the manner in which

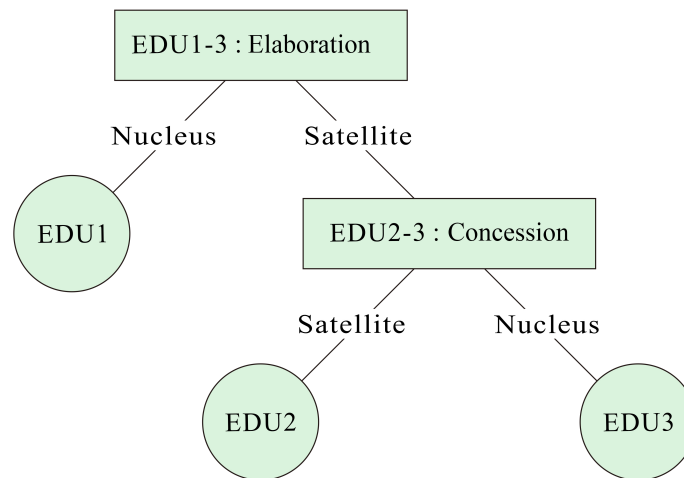


Figure 3.1: An example of RST tree: [*Rhetorical structure theory (RST) is a theory of text organization.*]^{EDU1} [*Although the RST structure is difficult to annotate,*]^{EDU2} [*there are still many scholars who have studied it.*]^{EDU3}

two text segments are interconnected (e.g., one segment might give a reason for a claim made in another segment, or alternatively, two segments may contrast with one another). Moreover, RST distinguishes between paratactic relations, where both segments carry equivalent discourse importance, and hypotactic relations, which classify the segment of greater centrality to the overarching discourse structure as the *nucleus* and the less central one as the *satellite*. Figure 3.1 shows a simple example of an RST tree. In this instance, EDU₁ serves as the nucleus of the elaboration relation, whereas the combination of EDU₂ and EDU₃ constitutes the satellite of said relation. Furthermore, we can see that EDU₃ assumes a more central role within the concession relation, hence it is marked as its nucleus, while EDU₂ is less important. If EDU₂ was left out, the elaboration relation between EDU₁ and EDU₃ would still hold, but if EDU₃ was removed, an elaboration relation between EDU₁ and EDU₂ would not hold, and the coherence would be lost. As has been recognized early on [Marcu. 1997](#); [Marcu. 1999b](#), this discourse information can be effectively used in summarization tasks.

While there have been some previous attempts at integrating discourse structure into neural text summarization models, as seen in [Gabriel et al. \(2021\)](#), [Dong et al. \(2021\)](#), [Xiao et al. \(2020\)](#), [Xu et al. \(2020\)](#), and [Cohan et al. \(2018\)](#), these approaches do not utilize relation labels and solely consider the 1-best RST tree obtained from preprocessing of a discourse parser. We argue that this leads to two significant issues. Firstly, infor-

mation pertaining to the relation type is overlooked, despite its known relevance to the summarization task. Secondly, there may be benefits in considering distributions over coherence relation labels, rather than limiting analysis to the 1-best results (Liu et al. 2022b). One reason is that external discourse parsers are known to perform poorly on out-of-domain data (Atwell et al. 2022; Liu et al. 2021e; Gessler et al. 2021; Koto et al. 2021b; Liu et al. 2020; Nguyen et al. 2021), and may hence propagate errors into the summarization model. There is a subsequent risk that these errors will be incrementally amplified during back-propagation, thus potentially impairing the model’s performance. A second reason is that there might inherently be several coherence relations holding at the same time (Yung et al. 2022), which might be beneficial to represent through the distributions of the discourse structure. Hence, we posit that the output of the RST parser holds greater significance when it not only provides the model with the n-best results but also conveys the remaining uncertainty associated with them. In the remainder of the chapter, we explore whether incorporating the labeled discourse relation structure with uncertainty, which can be understood as the distributions of discourse structure, into the attention mechanism can augment the performance of neural summarization models.

3.2 Methodology

In the realm of document discourse parsing, the performance of the RST parser leaves much to be desired (Yu et al. 2022; Nguyen et al. 2021; Liu et al. 2021e), with parsing performance deteriorating in conjunction with escalating document complexity. Merely passing the 1-best RST tree risks imparting misleading information to the summarization model. Inspired by Liu et al. (2022b), the approach to alleviating the aforementioned problems is that we retain uncertainty inside the parser, which can convey the parser’s confidence in each discourse relation. Furthermore, we contend that discourse relation labels (types) can provide more fine-grained labeled probability distributions that can assist attention heads of the Transformer-based model to capture the importance of different discourse units. This in turn would contribute to a more precise estimation of the context vector and can enhance the quality of source document encoding. Discourse parsers tend to be more precise (and have more peaked probability distributions) for local coherence relations, which span only a short amount of text, compared to global relations spanning large portions of a text. This aligns well with the dilated (yet still limited) sliding window attention mechanism of the Longformer (Beltagy et al. 2020).

We, therefore, integrate the probability distributions over local coherence relations into the attention window w of the Longformer.

RST Tensor with Labeled Distributions. The discourse-driven neural seq2seq summarization task can be modeled as follows:

$$P(t|s, d) \approx \prod_{i=1}^T P(t_i|t_{<i}, \text{encode}(s, d)) \quad (3.1)$$

In the above equation, s , t , and d denote the source, target sequence, and discourse representation, respectively. T signifies the target sequence length and $\text{encode}(\cdot)$ represents the encoder of the summarization model. Previous research (Xu et al. 2020; Cohan et al. 2018; Dong et al. 2021; Li et al. 2020c; Chen et al. 2021) has confirmed that the probability of generating appropriate summaries by incorporating d into the model’s encoder is greater than the probability of generating proper summaries without the incorporation of d . Therefore, our main idea is to find a better method to incorporate discourse structure d . To inject discourse structure, we first apply a matrixization approach to represent the discourse structure and produce a compact tensor representation appropriate for the Longformer model (Liu et al. 2022b).

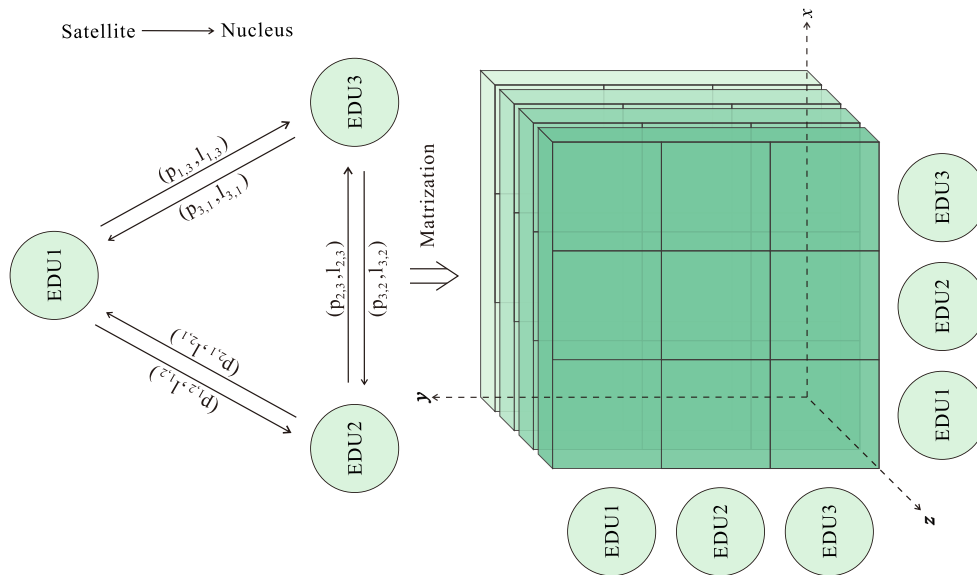


Figure 3.2: Labeled discourse distributions.

Figure 3.2 illustrates, by an example, how we convert the graph of all potential RST

relations (including the n-best RST trees) into a three-dimensional labeled discourse distribution (LDD) tensor. The x-axis and y-axis of the tensor represent the elementary discourse units (EDUs) in the source document, while the z-axis represents the type of discourse relation. Each point represents a confidence value $p(edu_i, edu_j, l) \in [0, 1] \subseteq \mathbb{R}$ of an elementary discourse unit edu_i connecting to another elementary discourse unit edu_j from the source text via the relation l . It should be noted that the generation of the LDD tensor should meet the conditions: 1) $p(edu_i, edu_i) = 0$, as no unit is dependent on itself; 2) we only extract the relation probability of nucleus units, since nucleus EDUs are more central to the text and should be given more attention.

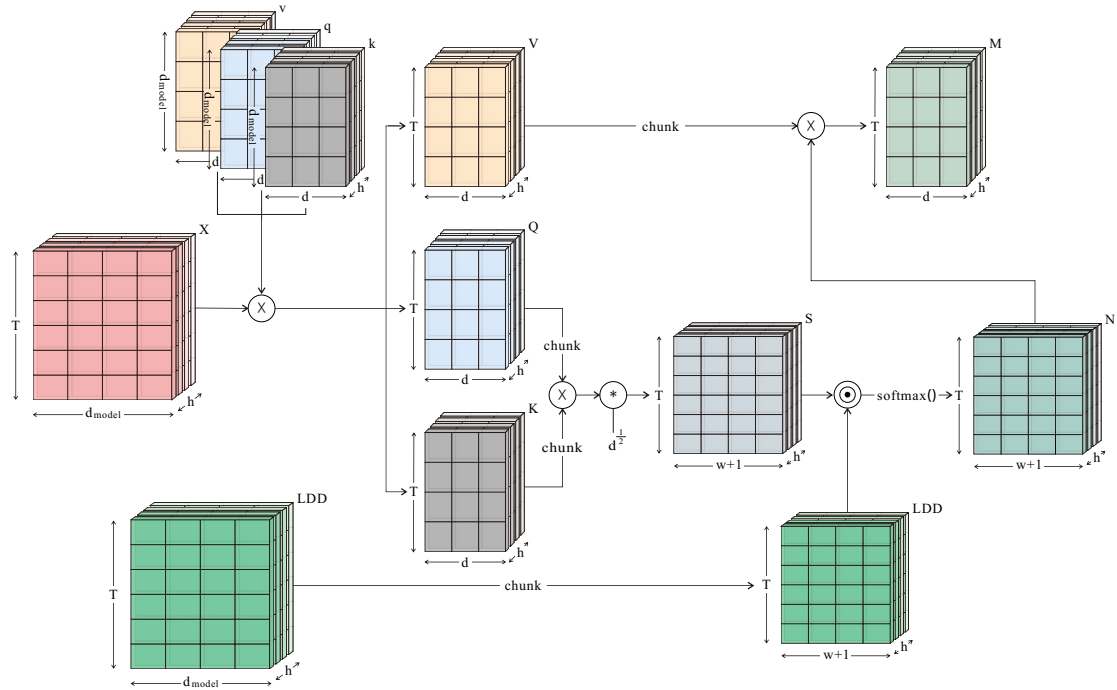


Figure 3.3: Model architecture. We show a schematic diagram of incorporating LDD tensor into the attention layer of the model. Specifically, X is the text embedding matrix, and LDD is incorporated with the attention matrix S in the form of element-wise multiplication. In order to ensure the consistency of matrix shape, we also apply an identical *chunk* method as Longformer in LDD.

RST-Injected Sparse Attention. We propose a novel Longformer-based summarization model, RSTformer, which incorporates LDD into each layer of the Longformer encoder in a discourse-aware manner. Given that each encoder layer shares an identical configuration, Figure 3.3 displays one layer architecture of our proposed model. The

standard dilated sliding window attention layer of Longformer employs a multi-head fixed-size window attention mechanism. For a pre-specified window size w , each token attends to $\frac{1}{2}w$ tokens on either side. For an input sequence of length T , the input of dilated sliding window attention heads in the RSTformer layer comprises the hidden representation tensor $X \in \mathbb{R}^{T \times d_{model} \times h}$ and labeled discourse distribution tensor $LDD \in \mathbb{R}^{T \times d_{model} \times h}$, where d_{model} represents the size of the hidden representation and h denotes the number of attention heads. As usual in multi-head self-attention, we multiply the text feature representation tensor with $q, k, v \in \mathbb{R}^{d_{model} \times d \times h}$ to obtain the corresponding $Q \in \mathbb{R}^{T \times d \times h}$, $K \in \mathbb{R}^{T \times d \times h}$, and $V \in \mathbb{R}^{T \times d \times h}$ matrices, where $d = d_{model}/h$. Subsequently, the attention weight matrix is obtained by:

$$S = \frac{Q \cdot K^\top}{\sqrt{d}} \quad (3.2)$$

Longformer utilizes two sets of projections, Q_s, K_s, V_s to compute the attention scores of sliding window attention, and Q_g, K_g, V_g to compute attention scores for global attention. Notably, Q_g, K_g, V_g are all initialized with values that match Q_s, K_s, V_s respectively. The dilated sliding window attention operates by calculating a fixed number of diagonals of QK^\top through sliding chunks query-key multiplication. This process yields a resulting tensor $S \in \mathbb{R}^{T \times w+1 \times h}$. Similarly, LDD and V adopt the same *chunk* method as employed by Longformer to acquire the sliding window attention matrix.

It should be noted here that we inject the sliding window attention tensor S obtained from the preceding computation by element-wise multiplication with the LDD tensor:

$$S \odot LDD \quad (3.3)$$

The motivation behind employing element-wise multiplication is to allow the learning parameters of the attention mechanism *dynamically* to optimize the summarization objective, but also diverge the least from the parser probabilities in discourse distribution (Liu et al. 2022b). The estimation of attention weights is adjusted to align with the utility of discourse relations for the ultimate summarization task. Following, the obtained weights are further processed using the softmax function to derive the final tensor representing the discourse-infused distribution:

$$N = \text{softmax}(S \odot LDD) \quad (3.4)$$

RST type	RST label
<i>Temporal</i>	Temporal
<i>Contingency</i>	Cause, Condition
<i>Comparison</i>	Comparison, Contrast, Concession, Topic-Change
<i>Expansion</i>	Explanation, Elaboration, Background, Topic-Comment

Table 3.1: RST relation category.

It should be emphasized that each attention head is assigned a different discourse matrix LDD_l for a specific relation l . This allocation enables heads to concentrate on and learn different discourse labels. In doing so, attention heads can be specialized and acquire a deeper understanding of the impact of discourse labels. The grouping of discourse relations is listed at [Table 3.1](#).

Finally, the discourse-injected weights N are multiplied by the value matrix V to obtain the attention weights M for this layer and then transfer M to the next Longformer encoder layer for further computation.

$$M = N \cdot V \quad (3.5)$$

3.3 Experimental Setup

Parser. We employ an external RST parser called DMRST ([Liu et al. 2021e](#); [Liu et al. 2020](#)) to automatically parse the source documents. The probability or uncertainty of discourse relations is extracted from the logits layer of the DMRST model.¹ In cases where DMRST fails to parse the source document, we simply skip the LDD generation process and proceed with the normal Longformer procedure.

Datasets. We conduct our experiments on three long document summarization datasets: BookSum Chapter ([Kryscinski et al. 2022](#)), eLife ([Goldsack et al. 2022](#)), and Multi-LexSum ([Shen et al. 2022](#)). We choose these datasets because of their high heterogeneity, and we want to investigate whether our approach can maintain adequate generalization performance across different data domains. [Table 3.2](#) shows the statistics of the datasets.

¹ https://github.com/seq-to-mind/DMRST_Parser

Dataset	Training	Validation	Test	Avg. Doc Words	Avg. Summary Words	Coverage	Density	Compression Ratio
BookSum Chapter	9600	1431	1484	3834.40	363.81	0.764	1.504	15.198
eLife	4346	241	241	10133.07	382.69	0.819	1.761	27.650
Multi-LexSum	3177	454	908	58210.99	547.04	0.926	3.394	95.390

Table 3.2: Datasets statistics.

In [Table 3.2](#), *Coverage* refers to the percentage of words in the summary that are from the source document. A higher coverage ratio indicates that a greater proportion of summary words are derived directly from the source text. It is mainly used to measure the degree of derivation of the summary from the text. *Density* is defined as the average length of the extracted segments to which each summary word belongs ([Segarra Soriano et al. 2022](#)). *Compression ratio* is defined as the ratio between the length of the source document and the summary ([Scialom et al. 2020](#)).

Evaluation Metrics. We evaluate the quality of different summarization systems using Rouge- $\{1, 2, L\}$ score ([Lin. 2004](#)), BERTscore ([Zhang et al. 2020b](#)), Meteor score ([Banerjee et al. 2005](#)), $\{1, 2, 3, 4\}$ -gram novelty, SummaC ([Laban et al. 2022](#)), and sentence alignment ([Liu et al. 2021c](#)) as criteria for the model’s effectiveness.

In detail, Rouge- $\{1,2\}$ is evaluated based on the co-occurrence of $\{1,2\}$ -gram in summary, while the calculation of Rouge-L uses the longest common subsequence. BERTScore is used to compute the semantic similarity score of candidate sentences to reference sentences through contextual embedding. Meteor is an improvement based on BLEU ([Papineni et al. 2002](#)), which also considers the impact of sentence fluency and synonyms on semantics. $\{1, 2, 3, 4\}$ -gram novelty indicates the capacity of the model to generate new words, rather than merely extracting words from the original text. SummaC detects semantic inconsistency by segmenting documents into sentence units and aggregating scores between sentence pairs.

Training and Inference. Hyperparameters for the baseline, proposal models, and ablation models are all kept identical. We adopt the same configuration as Longformer ([Beltagy et al. 2020](#)). All experiments are optimized using Adam ([Kingma et al. 2015](#)) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$, and weight decay = 0.1) with Adafactor ([Shazeer et al. 2018](#)), the number of warm-up steps is 1500, and the initial learning rate is set to $3e^{-9}$ with cosine learning rate schedule. We also apply NoisyTune (Noise lambda = 0.2) ([Wu et al. 2022](#)) for efficient fine-tuning. The size of the local attention window is $w = 1024$,

and we choose cross-entropy as the loss function.

During the training phase, we save the checkpoint with the highest Rouge-2 F1 score on the validation set as the final model. The experiments are all run for 30 epochs using a batch size of 1 with early stopping implemented. In order to prevent over-fitting, we set the dropout rate to 0.1 in all layers of the model. For model inference, we adopt a beam size of 4 with a length penalty of 2.0, and we set the no-repeat n-gram size to 3.

3.4 Results and Analysis

General Results. The experimental results for each model are presented in [Table 3.3](#). To estimate a lower bound in performance, we simply use the original document as the summary. Further trivial models include the Lead-3 model, which simply picks the first three sentences of the document as the summary. Lead-K similarly extracts the first K sentences of the document, until a similar length as the reference summary is reached. Longformer and state-of-the-art (SOTA) models serve as our baseline and comparison models, respectively. The remaining two models are the models we proposed. RSTformer (w/o relations) refers to the model that preserves whether there are relations between EDUs and ignores the type of relations by summing the third dimension of LDD tensors. RSTformer (w/ relations) is the final model we propose, with the only difference being the inclusion of the impact of RST types.

Both RSTformer versions are found to outperform the baseline model on various measures. The higher scores reflect an improved choice of words (Rouge & Meteor scores), and also the semantics of the text (BERTscore).² The proposed model, RSTformer, demonstrates robust generalization capabilities across different datasets, highlighting its potential in different summarization domains.

In most of our summarization experiments, we furthermore find that incorporating discourse structure with types provides better experimental results than the discourse distributions without types, even beating the SOTA model on our experimental datasets. This observation suggests that providing more discourse information, especially type distribution probabilities, is a promising approach.

² The version of BERTscore we use comes from the original paper version (Zhang et al. 2020b) with HuggingFace default API (<https://huggingface.co/spaces/evaluate-metric/bertscore>).

Dataset	Model	Rouge-1 F1	Rouge-2 F1	Rouge-L F1	BERTscore	Meteor
BookSum Chapter	Full article (lower bound)	13.742	4.019	13.421	0.805	21.299
	Lead-3	17.683	2.747	16.708	0.812	9.815
	Lead-K	29.149	4.641	28.034	0.805	24.091
	Longformer(baseline)	33.636	9.626	32.611	0.846	27.160
	RSTformer(w/o relations)	33.604	10.149	32.631	0.850	26.811
	RSTformer(w/ relations)	34.019	10.275 ^{†‡}	32.870	0.853 ^{†‡}	27.473 [‡]
	SOTA model (Kryscinski et al. 2022)	37.510	8.490	17.050	0.156	-
Our compared to baseline		+ Δ 0.383	+ Δ 0.649	+ Δ 0.259	+ Δ 0.007	+ Δ 0.313
Our compared to SOTA		- Φ 3.491	+ Φ 1.785	+ Φ 15.820	+ Φ 0.697	Φ -
eLife	Full article (lower bound)	6.893	2.327	6.675	0.831	13.864
	Lead-3	16.266	3.634	15.088	0.832	7.163
	Lead-K	37.188	7.971	35.151	0.832	25.331
	Longformer(baseline)	46.778	13.318	44.317	0.855	27.921
	RSTformer(w/o relations)	46.862	14.008	44.458	0.855	27.685
	RSTformer(w/ relations)	48.696 ^{†‡}	14.843 ^{†‡}	46.129 ^{†‡}	0.847	29.526 ^{†‡}
	SOTA model (Goldsack et al. 2022)	46.570	11.650	43.700	-	-
Our compared to baseline		+ Δ 1.918	+ Δ 1.525	+ Δ 1.812	- Δ 0.008	+ Δ 1.605
Our compared to SOTA		+ Φ 2.126	+ Φ 3.193	+ Φ 2.429	Φ -	Φ -
Multi-LexSum	Full article (lower bound)	3.862	2.198	3.786	0.784	8.825
	Lead-3	16.135	6.387	15.421	0.770	9.538
	Lead-k	29.145	9.276	27.734	0.784	24.266
	Longformer(baseline)	45.751	21.272	43.131	0.865	33.282
	RSTformer(w/o relations)	46.424	22.730	43.978	0.867	33.808
	RSTformer(w/ relations)	46.421	22.888 ^{†‡}	43.979	0.867 [†]	33.941
	SOTA model (Shen et al. 2022)	53.730	27.320	30.890	0.420	-
Our compared to baseline		+ Δ 0.670	+ Δ 1.616	+ Δ 0.848	+ Δ 0.002	+ Δ 0.659
Our compared to SOTA		- Φ 7.309	- Φ 4.432	+ Φ 13.089	+ Φ 0.447	Φ -

Table 3.3: Model performance. The bold numbers represent the best results with respect to the given test set. Δ and Φ represent the improvement of our model compared to the baseline and SOTA models, respectively. \dagger and \ddagger indicate statistical significance ($p < 0.05$) against the baseline model via T-test and Kolmogorov-Smirnov test. Each result of the three distinct SOTA models is directly replicated from their original papers.

Ablation Study. We also define two additional control conditions to examine the impact of RST attention (LDD) on model performance:

- **Without Attention Calculation (WAC):** We skip the previous calculation of attention weights and directly replace attention weights with the LDD tensor.
- **Random Identical Attention (RIA):** We assign fixed random values to the LDD tensor, regardless of the probability of discourse relations.

Table 3.4 shows that the RST attention cannot fully replace the calculation of the attention mechanism. Although the performance is lower than the baseline model, its

Dataset	Model	Rouge-1	Rouge-2	Rouge-L
BookSum	Longformer	33.636	9.626	32.611
	RSTformer(WAC)	31.956	8.772	31.049
Chapter	RSTformer(RIA)	32.881	9.067	31.899
eLife	Longformer	46.778	13.318	44.317
	RSTformer(WAC)	39.076	8.461	37.114
	RSTformer(RIA)	41.761	10.901	40.062
Multi-LexSum	Longformer	45.751	21.272	43.131
	RSTformer(WAC)	42.903	18.440	40.773
	RSTformer(RIA)	42.213	20.785	31.219

Table 3.4: F1 scores for ablation study.

main noteworthy advantage is that it saves considerable computations and parameters. Experiments by introducing random noise demonstrate that random values do indeed negatively impact the model’s performance. Furthermore, it also confirms the effectiveness of incorporating the probability distributions of discourse structure.

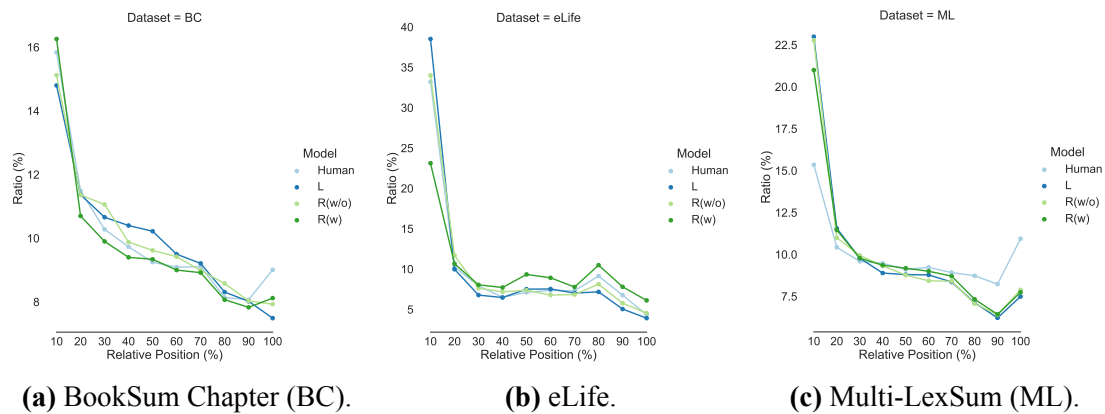
Human Evaluation. To better analyze the effectiveness of our model, we randomly select 10 samples from the BookSum dataset and hire human annotators to conduct the human evaluation. The recruited annotators are all master’s students or doctoral students with computer science-related or computational linguistics-related backgrounds. All annotators are compensated with the standard hourly salary set by the university. At the time of evaluation, we provide 3 candidate summaries for each source document, namely outputs from our final proposed model and baseline model, along with the ground truth summary. Each instance is assigned to 3 participants who are instructed to rate the faithfulness, informativeness, readability, and conciseness of the candidate summaries on a scale of 1 to 5. They are also supposed to give an overall rank of three summaries and identify which one is generated by humans. We provide the description of each criterion in our human evaluation in Appendix [Section A.1](#).

[Table 3.5](#) reports the human evaluation results. For each human evaluation indicator, we compute the average value to represent whether the candidate system has good performance in that indicator. Best and Worst indicate the proportion of times a summary by a particular model is judged to be best or worst among the three options. While neural summarization models still exhibit a notable performance gap when compared to human-generated summaries, our proposed model consistently outperforms the baseline model across all metrics.

Candidate	Faithful	Informative	Readable	Concise	Best Worst
Human	4.40	4.83	4.83	4.33	83.3% 0.0%
Longformer	2.50	2.57	3.43	2.70	6.7% 56.7%
RSTformer(w/ relations)	2.97	2.90	3.73	3.00	10.0% 43.7%

Table 3.5: Human evaluation results.

Sentence Alignment. We examine the alignment distributions of generated summaries to explore whether the improved model can be closer to human-summarized text (Liu et al. 2021c). Our results are depicted in Figure 3.4a, Figure 3.4b and Figure 3.4c. From a broader perspective, the sentence alignment distribution of our proposed models is more closely aligned with that of human summarizers. In addition, the generated summaries produced by our models demonstrate a greater emphasis on the content of the second half of the document, resulting in summaries that are more comprehensive and coherent in nature.

**Figure 3.4:** Sentence alignment distributions. L = Longformer, R(w/o) = RSTformer(w/o relations), R(w) = RSTformer(w/ relations).

N-gram Novelty & Inconsistency Detection. We also study the level of abstractive-ness and factual consistency in the generated summaries. To evaluate the abstractive-ness, we employed N-gram novelty as a measure to determine whether the model can generate words that are not present in the original text, rather than solely extracting content from the source document. For inconsistency detection, we utilize the SummaC method (Laban et al. 2022) for testing. Our results are shown in Figure 3.5 and Figure 3.6, respectively.

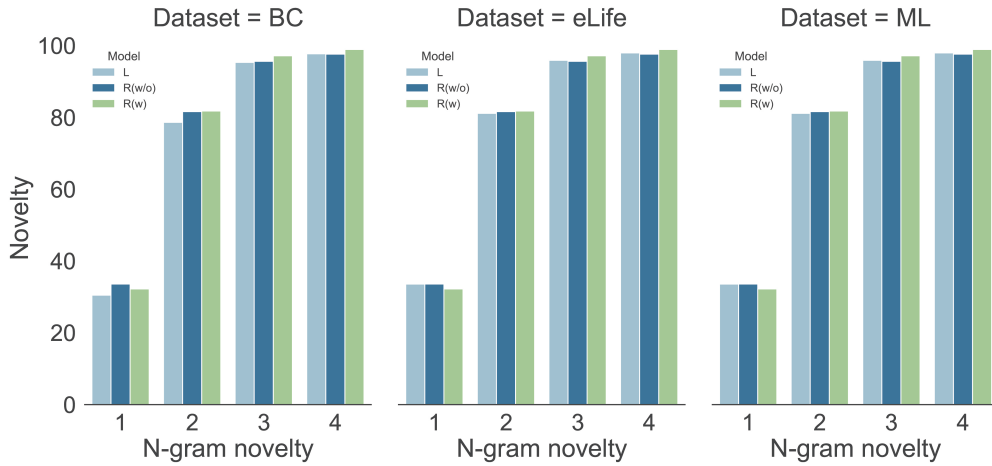


Figure 3.5: N-gram novelty. L = Longformer, R(w/o) = RSTformer(w/o relations), R(w) = RSTformer(w/ relations), BC = Booksum Chapter, ML = Multi-LexSum.

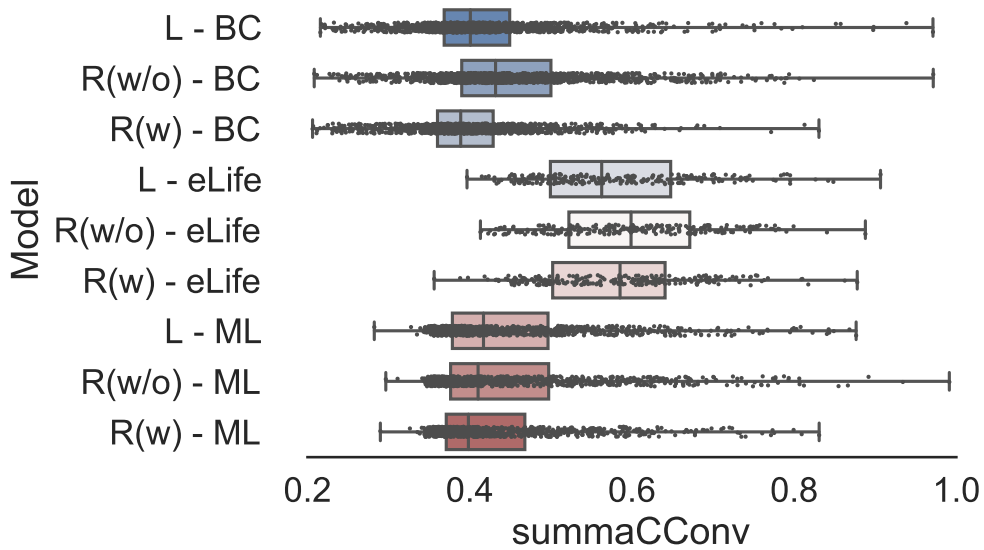


Figure 3.6: Consistency check. L = Longformer, R(w/o) = RSTformer(w/o relations), R(w) = RSTformer(w/ relations), BC = Booksum Chapter, ML = Multi-LexSum.

Compared with the baseline model, incorporating discourse information into the model does increase the ability of the model to generate novel words, especially evident in the context of 3-gram and 4-gram, the gap becomes more prominent. In addition, the proposed model also performs better than the baseline model in terms of model consistency checks. Due to the increased ability to generate creative words, the semantic

coherence ability of the models incorporating typed discourse relations is lower than that of models without typed discourse relations.

3.5 Related Work

Text Summarization with RST. Rhetorical Structure Theory offers a paradigm for describing how various discourse units relate to one another in a text. The RST tree structure, as illustrated in [Marcu \(1997\)](#) and [Louis et al. \(2010\)](#), can serve as a valuable tool for content selection in the process of summarization. For instance, [Kikuchi et al. \(2014\)](#) characterized the dependencies between sentences by constructing RST trees and pruning the parts that are marked as *satellites* while preserving the important content (*nucleus*) of the document as predicted summaries. Although RNN-based models are sometimes argued to be sufficient in implicitly learning discourse and semantic relations, [Liu et al. \(2019a\)](#)'s work underscored the value of explicitly integrating RST trees into the summarization model, thereby highlighting the significance of discourse relations for the neural summarization network. It is also worth noting that while the attention mechanism can more effectively uncover discourse relations without explicit training, it tends to unearth only superficial discourse structure and is often prone to mistakes ([Vig et al. 2019](#); [Sachan et al. 2021](#); [Xiao et al. 2021](#); [Huber et al. 2022](#); [Davis et al. 2020](#)).

Although attention-based models excel in executing downstream tasks such as summarization, the explicit incorporation of discourse relations can yield additional benefits. Work highly related to ours includes the model of [Xiao et al. \(2020\)](#), which improved the performance of an extractive summarization model by transmuting the RST structure into a dependency tree and explicitly integrating it into the computation of the attention mechanism. Follow-up works [Xu et al. \(2020\)](#) and [Dong et al. \(2021\)](#) further confirmed the influence of RST structure on improving attention mechanism by incorporating discourse structure into a transformer-based model and a graph neural network model for the summarization task, respectively. However, all of these neural strategies apply the one-best structure derived from an external discourse parser.

Text Summarization with Longformer. The Longformer model ([Beltagy et al. 2020](#)), based on a sparse attention mechanism, is considered to be an effective means for processing long documents. Its essence is to make each token only pay attention to a window of a certain size, so that the time complexity of the model is reduced from a quadratic

correlation with the text length to a linear correlation. Longformer-related models have since been employed in several summarization tasks (e.g., Zhang et al. 2022; Otmakhova et al. 2022; Elaraby et al. 2022; Xie et al. 2022; Liu et al. 2022a).

At the same time, there have also been recent attempts at integrating text structure information with the Longformer model in summarization tasks. Huang et al. (2021) first employed the Longformer to encode input documents and propose an extractive summarization model based on a heterogeneous graph of discourse and coreference relations. Liu et al. (2021b) extended the Longformer to model different types of semantic nodes in the original text as heterogeneous graphs and directly learn relations between nodes. Specifically, they treated tokens, entities, and sentences as different types of nodes, and the multiple sparse masks as different types of edges to represent relations (e.g., token-to-token, token-to-sentence). Elaraby et al. (2022) improved the performance of the strong baseline Longformer by integrating argument role labeling into the summarization process to capture the argumentative structure of legal documents. Ruan et al. (2022) and Cao et al. (2022b) enhanced extractive and abstractive summarization tasks, respectively, by introducing the text’s hierarchical structure (e.g., section title) into the Longformer model.

3.6 Chapter Summary

This chapter addressed Research Question 1 by investigating how discourse structure can be transformed into an uncertainty-aware signal that directly participates in content selection and ordering for long-document summarization. The proposed RSTformer model integrates labeled discourse distributions into the sliding-window attention of Longformer, allowing the encoder to modulate attention weights according to both the type and the confidence of rhetorical relations between elementary discourse units. Experiments on BookSum Chapter, eLife, and Multi-LexSum showed that retaining relation types and parsing uncertainty together tends to yield more consistent improvements than using either in isolation, and that the distributional representation is less susceptible to error propagation from the external parser than a single deterministic tree.

These findings offer support for Hypothesis I and suggest that distributional discourse representations can serve as a viable soft guidance mechanism for long-document encoding. At the same time, the experiments in this chapter assumed a fully trainable encoder, where all model parameters could be updated to absorb the discourse signal.

As model sizes and input lengths continue to grow, this assumption becomes increasingly impractical. The next chapter examines whether the same type of discourse guidance remains effective when the update budget is severely constrained, as is the case in parameter-efficient fine-tuning.

4

Discourse-Aware LoRA for Text Summarization

Chapter 3 establishes that discourse structure, when represented as uncertainty-aware distributions rather than deterministic parses, can better guide content selection and organization in long-document summarization. The preceding chapter extends RST parsing outputs from a single 1-best tree to a type-aware probabilistic distribution and injects this distributed information into the local attention computations of Longformer, which leads to consistent improvements across multiple datasets. These results provide empirical support for Research Question 1 (RQ1) and Hypothesis I and confirm that the representation of discourse relations in a distributed form proves more conducive to model absorption under parsing noise and structural ambiguity. However, the method proposed in **Chapter 3** assumes a relatively unconstrained training regime, where the model can freely update its full set of parameters. As model sizes and input lengths increase, this assumption becomes less viable due to the rapidly rising GPU memory and computational costs associated with long-sequence training. In practice, a common strategy is to freeze most model parameters and update only a small subset incrementally via parameter-efficient fine-tuning (PEFT). This approach achieves a more tractable trade-off between training feasibility and deployment cost. Against this backdrop, Research Question 2 (RQ2) asks how discourse signals can reliably guide content selection and organization under such constrained adaptation regimes with limited update capacity and computational resources.

To this end, **Chapter 4** tests Hypothesis II that, when integrated as structured and

uncertainty-aware guidance, discourse signals can exert an influence on content selection and inter-sentential organization even under constrained parameter update budgets. Concretely, we focus on the Low-Rank Adaptation (LoRA) framework, where weight updates are approximated by low-rank matrices. Prior work suggests that PEFT tends to improve surface-level fluency but struggles to consistently shape sentence organization or content selection, partly because discourse relations are not explicitly visible in the input as supervised labels. To address this issue, [Chapter 4](#) investigates whether RST-based discourse knowledge can provide *soft guidance* to the LoRA training process without increasing the number of trainable parameters. Instead of a binary choice regarding the presence of discourse structure, we decompose the discourse signal into four controllable injection forms along two axes, namely binary vs. probabilistic values (b vs. p) and coarse-grained (structure-only, wo) vs. fine-grained (relation-aware, w) granularity. The n -best discourse forests and their associated confidence scores are transformed into compact discourse distribution matrices, which serve as modulation coefficients applied via element-wise multiplication to LoRA’s input representations. Subsequent experiments on Longformer and Vicuna backbones across the Multi-LexSum, eLife, and BookSum datasets evaluate whether retaining uncertainty and discourse relation types can still yield measurable gains under PEFT constraints. [Chapter 4](#) also compares variants that incorporate only structural information with those that include relation-type distributions and analyzes their relative impact on summarization quality and factual consistency.

4.1 Problem Formulation

The advent of pre-trained large language models (LLMs), such as LLaMA ([Touvron et al. 2023](#)), Vicuna ([Zheng et al. 2023](#)), and GPT-related models from OpenAI ([OpenAI. 2023](#)), has greatly accelerated research progress in Natural Language Processing (NLP). With the continual growth in the scale of LLMs, the requirements for both software and hardware in order to fully fine-tune LLMs to adapt to downstream tasks, especially in processing long sequence data, will become increasingly demanding ([Gu et al. 2022](#); [Liu et al. 2024b](#)).

Parameter-Efficient Fine-Tuning (PEFT) strategies are noteworthy in mitigating the aforementioned problem by reducing the number of parameters that need to be adjusted ([Chen et al. 2022](#); [AkbarTajari et al. 2022](#); [Mao et al. 2022a](#); [Gheini et al. 2023](#); [Badola et](#)

al. 2023; Zhang et al. 2023c; Lawton et al. 2023). Some studies have highlighted that by updating only 0.01–1% of the (additional) parameters and freezing all other parameters of LLMs, PEFT methods can match or even exceed the performance of vanilla full-parameter fine-tuning (Li et al. 2021; Hu et al. 2022; Asai et al. 2022; Yang et al. 2022; Gu et al. 2023; Liao et al. 2023; Zhang et al. 2023c; Li et al. 2023; Lei et al. 2023; Zhang et al. 2023b; Chen et al. 2023; Lawton et al. 2023). Among these methods, the LoRA algorithm (Low-Rank Adaptation, Hu et al. 2022) has achieved state-of-the-art (SOTA) performance due to its ability to circumvent the latency associated with adapter tuning (Houlsby et al. 2019) as well as the input length constraints of prefix/prompt tuning (Li et al. 2021; Lester et al. 2021) during model training and inference (He et al. 2022; Ghazvininejad et al. 2022; Dettmers et al. 2023; Zhang et al. 2023b; Whitehouse et al. 2024; Ding et al. 2023).

Recent investigations (Üstün et al. 2022; Ponti et al. 2023; Zhao et al. 2023; Zeng et al. 2023; Zhang et al. 2023c; Wan et al. 2023; Liu et al. 2023e) have revealed that PEFT strategies face challenges in distinguishing latent text relations and determining the importance level of different sentences during fine-tuning. This issue arises because such distinctions are not a primary focus in PEFT’s learning process and are not explicitly represented in the input data. However, this is essential for the task of long document summarization since generating a good summary often requires natural language generation (NLG) models to have the ability to discern salient information within the text and comprehend the intricate interrelations among different text components.

Our approach proposed here takes inspiration from Liu et al. (2023c), who have advised that explicitly integrating document structure and/or discourse knowledge can enhance the performance of neural summarization models when fully fine-tuning the NLG models. This motivates us to investigate the following research questions: *Can the Rhetorical Structure Theory (RST, Mann et al. 1987) improve the performance of LoRA strategy in summarizing lengthy documents?* Specifically, we want to explore and verify whether infusing RST knowledge into LoRA can improve the performance of long document summarization. To answer this question, this chapter will propose, introduce, and integrate four RST structure variants to guide the training of LoRA. These variants include (i) binary (ii) probability RST distribution, both with (w/) and without(w/o) relation labels.

4.2 Methodology

A limitation observed with LoRA and other PEFT methods during the fine-tuning phase is that, while their primary function is to act as a low-rank approximation of the weight matrices in LLMs, they do not adequately capture textual context relations (He et al. 2022; Ghazvininejad et al. 2022; Wan et al. 2023; Zhang et al. 2023c). One of the reasons is that LoRA is not driven or guided by discourse knowledge during the training phase, because this part of knowledge is not explicitly present in the input data (Üstün et al. 2022; Zhao et al. 2023). In addition, the matrices obtained by low-rank approximation strategies may have more difficulty in capturing complex textual discourse relations due to the smaller semantic space that can be expressed when compared to LLMs' weight matrices (Wan et al. 2023; Zhang et al. 2023c; Tomanek et al. 2021). Hence, we propose a method that directly and explicitly incorporates discourse architecture into LoRA. This approach allows LoRA's adaptation mechanism for downstream tasks to discern intricate text relations through soft guidance, which can leverage contextual discourse connections and steer the learning trajectory toward a discourse-informed summarization process.

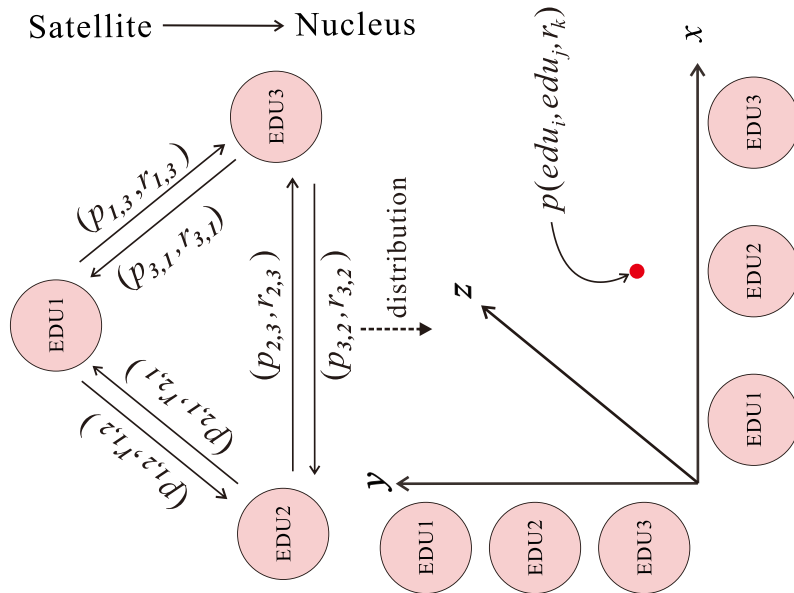


Figure 4.1: RST distribution.

RST Distribution. Our approach builds upon prior practice (Chen et al. 2017; Xiao et al. 2020; Bugliarello et al. 2020; Liu et al. 2022b; Liu et al. 2023c; Zhao et al. 2023) of integrating linguistic structures (such as syntactic structure, discourse structure, etc.) into neural NLG models. To infuse discourse structure into LoRA, we begin by converting RST structures, generated by an RST parser, into a compact matrix representation.¹ Figure 4.1 exemplifies how to transmute the full potential RST structures (n-best RST forests) into a three-dimensional discourse matrix (Liu et al. 2023c). In this matrix, the x and y axes correspond to the elementary discourse units (EDUs) within the source document, while the z -axis denotes the discourse relation label. Each point of the matrix indicates the probability value $p(edu_i, edu_j, r_k) \in [0, 1] \subseteq \mathbb{R}$ that edu_i is the nucleus of edu_j with discourse relation r_k . It should be noted that $\forall i = j, p(edu_i, edu_j, r_k) = 0$, since no unit is self-dependent. Next, we average and merge the y -axis of the matrix, and the merged value $c(edu_i, \overline{edu_j}, r_k)$ is called the importance index of edu_i with relation r_k . The RST distribution is then obtained by combining all $c(edu_i, \overline{edu_j}, r_k)$. Based on this, we propose four fine-grained RST matrix distributions:

- RST_{wo}^b : A binary, label-agnostic representation collapsing probabilities into a simple 1-or-0 regarding discourse connections.
- RST_w^b : An extension of the binary distribution that includes relation labels, enriching the binary decisions with relational types.
- RST_{wo}^p : A probabilistic representation that omits labels, focusing instead on the probabilities to express uncertainty in discourse connections.
- RST_w^p : The most granular representation, retaining both types of discourse relations and their probabilistic weights for a full-fledged representation of discourse nuances.

The inclusion of the relation label is contingent on whether we perform average-and-merge along the relation dimension (z -axis). Whether the approach is binary or based on uncertainty hinges on whether we replace the probability value with 1 or 0. In the binary cases, probabilities equal to or above 0.5 are replaced with 1, else with 0. Previous researchers (such as Xu et al. (2020) and Dong et al. (2021)) considered the 1-best tree, representing binary relations outputted from parsers into summarization models (also the case of our first two variants). The latter two variants utilize the parser’s output probabilities as confidence indicators for discourse connections (Liu et al. 2023c).

¹ https://github.com/seq-to-mind/DMRST_Parser

RST-Aware Injection. In the process of vanilla LoRA fine-tuning, let $W_{A \times B}^{fine-tuned}$ denote the fine-tuned LLM’s parameters, and $W_{A \times B}^{pre-trained}$ represent the parameters before fine-tuning. The change in parameters is represented by $\Delta W_{A \times B}$, where A and B correspond to the dimensions of the parameter matrix:

$$W_{A \times B}^{fine-tuned} = W_{A \times B}^{pre-trained} + \Delta W_{A \times B} \quad (4.1)$$

In other words, the parameters after fine-tuning can be obtained by adding a matrix representing the variation to the parameters of the original, pre-fine-tuned model.

$$\Delta W_{A \times B} \simeq \Phi[(W_{A \times r}^{down} W_{r \times B}^{up})] \quad (4.2)$$

$$r \ll \min(A, B) \quad (4.3)$$

The objective of the LoRA strategy is to learn the mapping method Φ that can provide an approximation of the matrix representing parameter variations (Hu et al. 2022). Typically, the rank value r is considerably smaller than both A and B , so that the total number of parameters of $W_{A \times r}^{down}$ and $W_{r \times B}^{up}$ is significantly smaller than $W_{A \times B}$. For a given input document X to the linear projection in the model’s hidden layer, LoRA modifies the projection output (hidden representation) h as follows:

$$h \leftarrow h + X(W_{A \times r}^{down} W_{r \times B}^{up}) \quad (4.4)$$

In its current form, LoRA treats both satellite and nucleus EDUs in documents equally and only recognizes their difference during the back-propagation process. This issue is also noted in the analyses by Ghazvininejad et al. (2022) and Zhao et al. (2023), who also discovered that PEFT faces challenges in understanding the complex relations between sentences and the differences in importance level between text segments during its learning process. Therefore, we soft-guide the learning process by injecting the RST structure (i.e., the matrix presentation mentioned above) into the text embedding matrix of LoRA, as shown in Figure 4.2. Specifically:

$$h \leftarrow h + [(X \odot (1 + \gamma))(W_{A \times r}^{down} W_{r \times B}^{up})] \quad (4.5)$$

Here, γ denotes the weight coefficient matrix, or more precisely, the RST distribution matrix. The operation \odot signifies element-wise multiplication, and the motivation

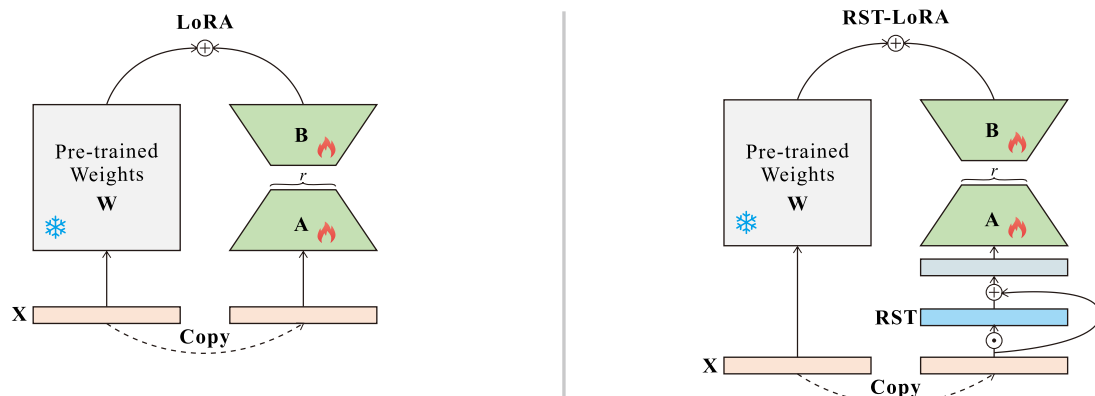


Figure 4.2: Model architecture. The diagram illustrates the integration of the RST matrix into the LoRA model. The left side is the original LoRA, while the right side depicts our proposed method RST-LoRA.

behind employing element-wise multiplication is that it can amplify the impact of probability values on the input X matrix, creating an RST-injected matrix with greater distributional variance; in contrast, element-wise addition would exert a lesser impact on X . It should be noted that the RST parser operates at the EDU level, meaning that sub-word units within the same EDU share the same multiplication factor, embedding the same probability value across the entire EDU into X . The estimates of learned parameters $W_{A \times r}^{down}$ and $W_{r \times B}^{up}$ are adjusted to match the utility of discourse knowledge for the ultimate summarization purpose. Each element of γ is constrained to be non-negative. The operation of $1 + \gamma$ functions as a residual connection, allowing discourse knowledge to exert a subtle influence on the adjustment of the low-rank weight matrix. If we set all elements of γ to a uniform value δ , including zero, the adjustment to the low-rank matrices would revert to the conventional LoRA approach.

4.3 Experimental Setup

Datasets. Our experiments are conducted on three long document summarization datasets: Multi-LexSum (ML, Shen et al. 2022), eLife (Goldsack et al. 2022), and BookSum Chapter (BC, Kryscinski et al. 2022).

Parser. For automatic parsing of source documents, we employ the DMRST parser (Liu et al. 2020; Liu et al. 2021e), which enables us to extract probabilities or uncertainties

of discourse relations and type labels from its final logits layer.

Automatic Metrics. Aligning with previous work for evaluating summarization systems (Narayan et al. 2018; Liu et al. 2023g; Blinova et al. 2023), we use F1 scores of Rouge-1 (R1), Rouge-2 (R2), Rouge-L (RL), and Rouge-Lsum (RLsum) (Lin. 2004), BERTScore (Zhang et al. 2020b), METEOR (Banerjee et al. 2005), sacreBLEU (Post. 2018), and NIST (Lin et al. 2003) for the model’s performance evaluation.

Training & Inference. We operate Longformer (Beltagy et al. 2020) and Vicuna13B-16k (Zheng et al. 2023) as our baseline backbone models. Longformer is a state-of-the-art, open-source model optimized for handling long documents under a Seq2Seq architecture. Meanwhile, Vicuna is another SOTA model based on GPT architecture. Our objective in using these models is to demonstrate the generalizability of our strategy across different architectural frameworks. We also include GPT-4 (OpenAI. 2023) as one of our comparative models. It should be noted that for GPT-4, we use both zero-shot learning (ZS) and in-context learning (ICL) with demonstrations from two randomly selected samples from the training datasets.² Besides, we compare our results with both the original full parameter fine-tuning (FFT) and the vanilla LoRA fine-tuning. All open-source models, including the baseline, proposed, and ablation models, adhere to identical hyperparameter settings.

Specifically, all experiments are optimized using the Adam (Kingma et al. 2015) optimizer (with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$, and weight decay = 0.1) and Adafactor (Shazeer et al. 2018), with a warm-up ratio of 0.2. The initial learning rate is set to $5e-5$, with a cosine learning rate schedule. Additionally, within the LoRA strategy, we set a constant rank r to 8, the scaling α to 32, and the dropout rate to 0.1. During training, we save checkpoints that achieve the highest Rouge-2 F1 score on the validation set as the final model. All experiments are run for 50 epochs with a batch size of 16, and early stopping is implemented to prevent over-fitting (all models converged before 50 epochs). For model inference, we employ a beam search of size 4 with a length penalty of 3.0 and set a no-repeat n-gram size of 3.

For GPT-4, we employ GPT-4 Turbo version (*gpt-4-1106-preview*), which is, at the time of experimentation (between 10 October 2023 and 15 December 2023), the best-

² Prompts can be found in Appendix Figure B.1 and Appendix Figure B.2.

performing publicly accessible version provided by OpenAI.³ For the hyper-parameter setting, we set temperature=1, top_p=1, frequency penalty=0.2, and presence penalty=0.2. The remaining hyperparameters are set to their default values as recommended by OpenAI.

4.4 Results and Analysis

General Results. The differences in performance of different RST variants are shown in [Table 4.1](#). Among our proposed RST-injected variants, models integrating discourse relation labels generally outperformed those without this integration. Similarly, models considering the uncertainty in discourse relations fare better than those disregarding it. This suggests that integrating parser uncertainty and coherence labels into the model improves the robustness of the model against potential misinformation to a certain extent when compared to the parser’s 1-best binary decisions.

³ <https://platform.openai.com/docs/models/>

Data	Model	R1	R2	RL	RLsum
Multi-LexSum	Longformer _{RST_{w_o}^b-LoRA}	45.82	21.32	23.81	43.40
	Longformer _{RST_w^b-LoRA}	46.02	21.34	23.87	43.39
	Longformer _{RST_{w_o}^p-LoRA}	46.21	21.54	24.09	43.37
	Longformer _{RST_w^p-LoRA}	46.33	21.86	24.11	43.58
	Vicuna _{RST_{w_o}^b-LoRA}	46.32	21.64	24.22	43.32
	Vicuna _{RST_w^b-LoRA}	47.33	22.70	24.25	43.31
	Vicuna _{RST_{w_o}^p-LoRA}	47.39	22.79	24.35	43.33
	Vicuna _{RST_w^p-LoRA}	47.45	23.19	24.39	44.02
eLife	Longformer _{RST_{w_o}^b-LoRA}	49.34	14.24	21.34	46.74
	Longformer _{RST_w^b-LoRA}	49.41	14.39	21.29	46.79
	Longformer _{RST_{w_o}^p-LoRA}	49.87	14.49	21.83	47.15
	Longformer _{RST_w^p-LoRA}	49.89	14.68	22.11	47.64
	Vicuna _{RST_{w_o}^b-LoRA}	48.73	14.68	21.89	47.11
	Vicuna _{RST_w^b-LoRA}	49.72	14.72	22.03	47.02
	Vicuna _{RST_{w_o}^p-LoRA}	49.87	14.79	22.21	48.10
	Vicuna _{RST_w^p-LoRA}	49.92	14.92	22.41	48.21
BookSum Chapter	Longformer _{RST_{w_o}^b-LoRA}	34.70	10.22	20.39	34.21
	Longformer _{RST_w^b-LoRA}	34.72	10.19	20.41	34.87
	Longformer _{RST_{w_o}^p-LoRA}	35.29	11.38	21.62	35.11
	Longformer _{RST_w^p-LoRA}	35.40	11.76	21.88	35.27
	Vicuna _{RST_{w_o}^b-LoRA}	37.28	12.35	22.13	38.33
	Vicuna _{RST_w^b-LoRA}	37.41	12.66	22.51	38.40
	Vicuna _{RST_{w_o}^p-LoRA}	37.87	13.10	22.77	39.69
	Vicuna _{RST_w^p-LoRA}	37.92	13.24	22.93	40.31

Table 4.1: Performance of different RST variants.

Dataset	Model	# Trainable Parameters	R1	R2	RL	RLsum	BERTscore	Meteor	sacreBLEU	NIST
Multi-LexSum	Longformer _{FFT}	0.44B	45.81	21.32	23.71	43.25	87.21	33.30	12.06	2.23
	Longformer _{LoRA}	1.13M	45.78	21.30	23.65	43.12	87.31	33.31	12.00	2.28
	Longformer _{RST_w^p-LoRA}	1.13M	46.33 ^{†‡}	21.86 ^{†‡}	24.11 ^{†‡}	43.58 ^{†‡}	92.01 ^{†‡}	34.55 ^{†‡}	13.11 ^{†‡}	3.21 ^{†‡}
	Vicuna _{FFT}	13B	46.40	21.88	24.15	43.28	90.02	33.19	13.56	3.32
	Vicuna _{LoRA}	6M	46.32	21.76	24.09	43.14	89.45	33.22	13.44	3.31
	Vicuna _{RST_w^p-LoRA}	6M	47.45 [‡]	23.19 ^{†‡}	24.39^{†‡}	44.02^{†‡}	93.89^{†‡}	35.31^{†‡}	14.02^{†‡}	4.11^{†‡}
	GPT-4 _{ZS}	-	38.74	13.39	18.26	37.67	60.91	24.24	7.43	1.55
	GPT-4 _{ICL}	-	42.14	15.27	20.37	40.12	71.32	28.14	10.22	1.90
	Liu et al. (2023c)	-	46.42	22.89	-	43.98	86.70	33.94	-	-
	Shen et al. (2022)	-	53.73	27.32	-	30.89	42.01	-	-	-
eLife	Longformer _{FFT}	0.44B	47.59	13.58	20.75	45.25	85.50	28.21	6.86	2.90
	Longformer _{LoRA}	1.13M	48.31	13.69	21.10	45.80	85.63	28.18	7.05	3.12
	Longformer _{RST_w^p-LoRA}	1.13M	49.89 ^{†‡}	14.68 ^{†‡}	22.11 ^{†‡}	47.64 ^{†‡}	87.64 ^{†‡}	31.23 ^{†‡}	7.78 ^{†‡}	3.79^{†‡}
	Vicuna _{FFT}	13B	48.32	14.06	21.31	45.57	85.71	30.28	7.00	2.91
	Vicuna _{LoRA}	6M	48.41	14.32	21.40	46.01	86.06	31.00	6.62	2.88
	Vicuna _{RST_w^p-LoRA}	6M	49.92^{†‡}	14.92^{†‡}	22.41^{†‡}	48.21^{†‡}	87.81^{†‡}	33.22^{†‡}	8.15^{†‡}	3.42 ^{†‡}
	GPT-4 _{ZS}	-	42.73	9.05	17.93	40.15	61.21	25.13	3.47	2.32
	GPT-4 _{ICL}	-	44.62	11.35	20.03	44.09	73.23	27.36	5.66	2.45
	Liu et al. (2023c)	-	48.70	14.84	-	46.13	84.70	29.53	-	-
	BookSum Chapter	Longformer _{FFT}	0.44B	34.68	10.02	20.35	33.71	81.02	27.30	3.32
Longformer _{LoRA}		1.13M	34.63	9.96	20.22	33.79	81.33	27.32	3.55	1.86
Longformer _{RST_w^p-LoRA}		1.13M	35.40 ^{†‡}	11.76 ^{†‡}	21.88 ^{†‡}	35.27 ^{†‡}	83.99 ^{†‡}	29.03 ^{†‡}	5.94^{†‡}	2.02 ^{†‡}
Vicuna _{FFT}		13B	37.21	12.38	22.07	38.21	82.31	28.01	3.45	1.70
Vicuna _{LoRA}		6M	37.30	12.26	21.84	38.23	82.23	27.83	3.34	1.68
Vicuna _{RST_w^p-LoRA}		6M	37.92 ^{†‡}	13.24^{†‡}	22.93^{†‡}	40.31^{†‡}	84.12 ^{†‡}	29.22^{†‡}	5.48 ^{†‡}	2.32^{†‡}
GPT-4 _{ZS}		-	35.25	7.46	17.52	34.23	58.56	26.50	3.36	1.54
GPT-4 _{ICL}		-	37.42	10.06	19.49	36.11	79.56	27.56	3.52	1.72
Liu et al. (2023c)		-	34.02	10.28	-	32.87	85.30	27.47	-	-
Cao et al. (2024)		-	41.11	10.63	-	40.20	-	-	-	-
Scirè et al. (2023)	-	42.13	10.53	16.75	-	-	-	-	-	

Table 4.2: Model performance. The bold numbers represent the best results concerning the given test set. [†] and [‡] indicate statistical significance ($p < 0.05$) of our final model (RST_w^p-LoRA) against the FFT and LoRA model via paired t-test based on the same backbone, respectively. FFT for full fine-tuning, ZS for zero-shot learning, and ICL for in-context learning. Each result of the SOTA models is directly replicated from their original papers.

Table 4.2 shows the performance differences between our final strategy (the best RST variant) and other comparative models. Specifically, GPT-4 exhibits the poorest overall performance, attributable to a lack of parameter tuning. The performance of the models based on Vicuna as the backbone is overall better than the models based on Longformer due to the larger number of parameters. Regarding parameter-efficient

settings, vanilla LoRA’s performance is marginally lower than FFT across most datasets, except eLife. However, LoRA achieves comparable results to FFT while only requiring adjustments of 0.25% of parameters for Longformer and 0.05% for Vicuna, highlighting LoRA’s efficiency.

We also observe consistent performance improvements in LoRA when integrating RST structure into its training process without increasing the number of fine-tunable parameters, and in most cases even exceeds the FFT model. Our final model RST_w^p -LoRA, integrates both discourse relation types and uncertainty into LoRA’s training, achieving the best experimental outcomes. It also defeats SOTA models (fully fine-tuned with complicated strategies) on some metrics, including the current most advanced model (Liu et al. 2023c) that incorporates RST structure to improve summarization performance.

Ablation Results. To further assess the impact of the RST matrix on model performance, we specify three additional control conditions:

- RST_{Even} : We assign values of 1 at even positions and 0 at odd positions in the RST matrix.
- RST_{Odd} : We assign values of 1 at odd positions and 0 at even positions in the RST matrix.
- RST_{Random} : We assign random values $\in [0, 1] \subseteq \mathbb{R}$ to the RST matrix without considering the probability of discourse relations.

Dataset	Model	R1	R2	RL	RLsum
ML	RST_{Even}	46.21	21.39	23.66	42.55
	RST_{Odd}	46.26	21.37	23.82	42.90
	RST_{Random}	46.30	21.73	24.07	43.10
eLife	RST_{Even}	47.10	14.28	20.86	45.33
	RST_{Odd}	47.04	14.20	20.98	45.31
	RST_{Random}	47.32	14.29	21.36	45.71
BC	RST_{Even}	37.09	12.20	21.75	38.06
	RST_{Odd}	37.01	12.18	21.72	38.10
	RST_{Random}	37.27	12.23	21.80	38.19

Table 4.3: F1 scores for ablation study.

In ablation experiments, we use Vicuna as the backbone for testing. The motivation

behind setting these three ablation conditions is to simulate the extreme scenario where the RST parser completely fails to deliver valuable discourse information. Table 4.3 indicates that different ablation integration strategies not only fail to enhance the model’s performance but even detract from it. Experiments by introducing random noise show that these arbitrary values reduce the model’s performance to a level marginally lower than the original LoRA. Furthermore, this also implies that when the RST parser fails to provide meaningful knowledge (as in the case of random noise), the impact of noise on the performance of the model is limited.

Hallucination Checking. We delve deeper into the level of factual consistency of the generated summaries, which we test using the SummaC method (Laban et al. 2022). The score of SummaC ranges from 0 to 1, and the higher the score, the better the consistency. The results of the assessment using Vicuna as the backbone are depicted in Figure 4.3. We observe that GPT-4 exhibits the weakest factual consistency, while the original LoRA also shows a comparatively lower level of factual accuracy than FFT. However, explicitly incorporating RST structure into LoRA mitigates the issue of hallucinations/inaccuracies in generated summaries, achieving better results than the FFT model.

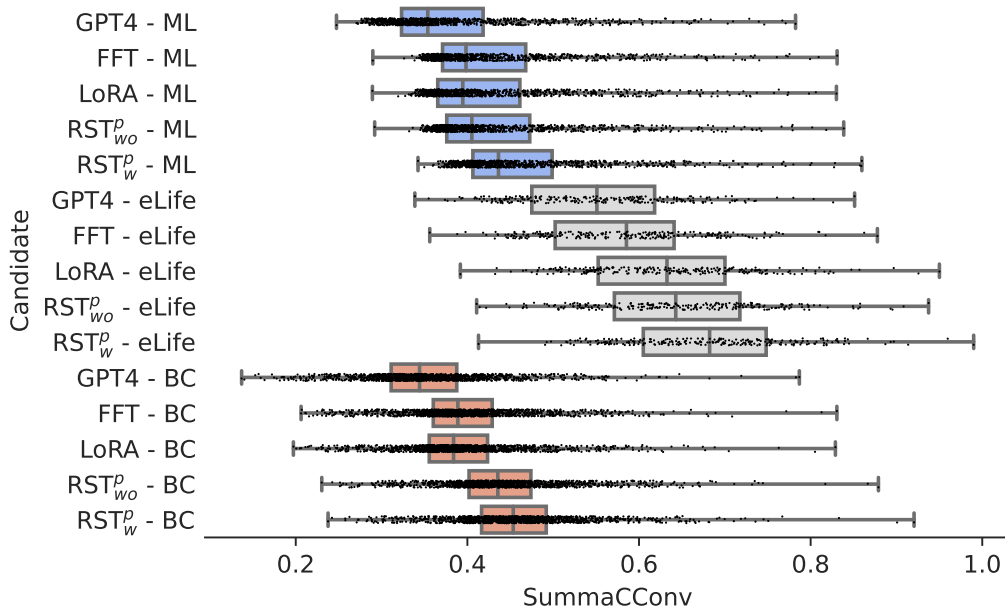


Figure 4.3: Factual consistency analysis.

Impact of Different Rank r . Figure 4.4a and Figure 4.4b and Figure 4.4c illustrate the impact of different ranks on model performance (Vicuna backbone). Across different datasets, the RST-aware model consistently outperforms the original LoRA at various ranks and achieves similar performance as the FFT model at lower ranks. Furthermore, a larger rank r will help to improve the performance of the model, which is also aligned with the findings of He et al. (2022) and Zhang et al. (2023b). However, a higher rank r correlates with an increased number of parameters requiring adjustment. Importantly, $r = 8$ is a trade-off point between performance gain and computational cost. When r continues to increase, the gain rate of performance improvement begins to slow down.

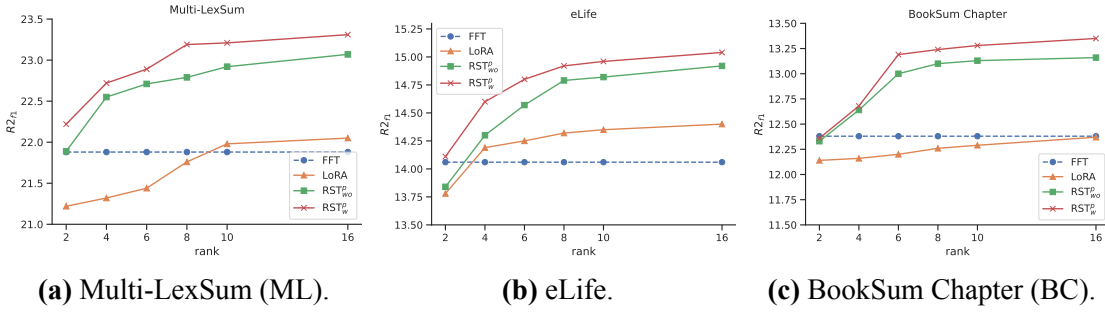


Figure 4.4: Impact of different r values on three datasets.

Impact of Parser Capability. To evaluate the parser’s impact on our method, we conduct an experiment that involves intentionally altering the RST parser’s output. This is designed to simulate varying levels of parser performance instability, thereby allowing us to observe its influence on our model’s efficacy. Specifically, we introduce random masking to the parser’s output at incremental thresholds of 10%, 20%, 40%, and 80%, assigning random values within the range of 0 to 1 to portions of the RST matrix. Table 4.4 presents the findings from this experiment, with Vicuna serving as the backbone for RST_w -LoRA model on the Multi-LexSum dataset.

These results illustrate the direct correlation between the RST parser’s performance and the performance of our model’s output. Notably, even under conditions of compromised parser performance (with up to 20% of the information being randomly masked), our model still demonstrates a good capacity to enhance summary generation quality by leveraging the learned discourse structure knowledge. However, it is observed that when the level of induced noise surpassed 40%, the negative impact became pronounced, relegating the model’s performance to levels akin to that of the original vanilla LoRA.

Model	R1	R2	RL	RLsum
RST_10%	47.33	23.01	24.33	43.45
RST_20%	47.09	22.78	24.23	43.37
RST_40%	46.52	21.76	24.13	43.20
RST_80%	46.32	21.75	24.06	43.15

Table 4.4: Impact of random masking on the parser.

Human Evaluation. To better analyze the quality of the summaries generated by the models, we randomly select 10 instances from the BookSum dataset and conduct a human evaluation. The evaluators we have recruited are graduate and doctoral candidates with specializations in Computer Science or Computational Linguistics, each possessing advanced proficiency in English. They receive compensation at the University’s established hourly rate. Evaluators are asked to read the corresponding original document, as well as five candidate summaries (from FFT, LoRA, and RST_w^p -LoRA with Vicuna backbone, GPT-4, and human). The human evaluators are blind to the condition, i.e., they do not know which summary comes from which system (or human author). Each sample is independently evaluated by three distinct human raters (thus, 150 evaluation samples in total). Evaluators should rate the candidate summaries on a scale of 1 to 5 for relevance, informativeness, conciseness, and faithfulness, with a higher score indicating better quality. They also need to give an overall ranking of the five summaries. The detailed guidelines for human evaluation are available in Appendix [Section B.2](#). The results, presented in [Table 4.5](#), show the average values for each metric, as well as the proportions of times each model’s output is considered the best or worst among the candidates. The scores of Fleiss’ Kappa coefficient for relevance, informativeness, conciseness, and faithfulness are 0.812, 0.705, 0.683, and 0.688, respectively, with an average score of 0.722, indicating substantial agreement.

From [Table 4.5](#), it is evident that human-generated summaries surpass all neural summarization models in terms of quality. Among the four neural models, GPT-4 shows the least performance, with LoRA coming in second, having a 20% probability of being rated as the worst. The FFT model fares slightly better than the LoRA model. The RST_w^p -LoRA model outperforms other neural summarization systems across all metrics, and its average scores on some indicators approach the level of human performance.

Candidate	Relevance	Informativeness	Conciseness	Faithfulness	Best Worst
Human	4.70	4.83	4.53	4.67	83.3% 0.0%
GPT-4 _{ICL}	3.76	2.27	3.25	2.33	0.0% 56.7%
Vicuna _{LoRA}	4.03	2.37	3.20	2.50	0.0% 20.0%
Vicuna _{FFT}	4.27	2.57	3.67	2.77	6.67% 13.3%
Vicuna _{RST_w^p-LoRA}	4.53	3.90	4.03	3.17	13.3% 10.0%

Table 4.5: Human evaluation results.

Moreover, compared to other neural summarization systems, the RST_w^p-LoRA model is more likely to be recognized for producing the highest quality summaries and less likely to be considered as generating the poorest quality summaries.

GPT-4 Evaluation. Inspired by Liu et al. (2023f), we engage GPT-4 to assess our candidate models using the same guidelines as our human evaluators. To ensure experimental consistency, all experiments use the identical hyperparameter settings. To avoid potential biases from previous interactions, we reset the conversation history prior to each query and abstain from making any further modifications. In our initial investigation, we aim to explore the extent to which GPT-4 evaluations generally concur with human assessments in terms of both relative ranking and average scores within the same subset of 10 samples delineated in human evaluations.⁴ We find through experiments that the above correlation does exist. We then extend the evaluation to include all samples from the test sets.⁵

Candidate	Relevance	Informativeness	Conciseness	Faithfulness	Best Worst
Human	4.89	4.76	4.67	4.72	96.8% 0.0%
GPT-4 _{ICL}	4.02	3.81	4.47	3.12	0.0% 35.3%
Vicuna _{LoRA}	4.20	3.82	4.43	3.37	0.0% 29.5%
Vicuna _{FFT}	4.31	4.04	4.49	3.55	0.0% 25.5%
Vicuna _{RST_w^p-LoRA}	4.46	4.44	4.60	4.12	3.2% 9.7%

Table 4.6: GPT-4 evaluation results on BC dataset.

⁴ Utilizing the same iteration of the GPT-4 model as employed in prior summary generation tasks.

⁵ Prompt can be found in Appendix Figure B.3.

Candidate	Relevance	Informativeness	Conciseness	Faithfulness	Best Worst
Human	4.67	4.70	4.52	4.83	94.2% 0.0%
GPT-4 _{ICL}	4.43	3.88	3.62	3.19	0.0% 43.3%
Vicuna _{LoRA}	4.52	4.03	4.20	3.40	0.0% 28.4%
Vicuna _{FFT}	4.52	4.06	4.28	3.58	0.0% 21.6%
Vicuna _{RST_w^p-LoRA}	4.57	4.33	4.31	4.22	5.8% 6.7%

Table 4.7: GPT-4 evaluation results on ML dataset.

Candidate	Relevance	Informativeness	Conciseness	Faithfulness	Best Worst
Human	4.80	4.81	4.72	4.78	96.3% 0.0%
GPT-4 _{ICL}	4.22	3.91	4.35	3.45	0.0% 45.2%
Vicuna _{LoRA}	4.47	4.12	4.41	3.58	0.0% 30.1%
Vicuna _{FFT}	4.59	4.23	4.47	3.82	0.2% 16.3%
Vicuna _{RST_w^p-LoRA}	4.62	4.49	4.63	4.39	3.5% 8.4%

Table 4.8: GPT-4 evaluation results on eLife dataset.

The outcomes for these tests are shown in [Table 4.6](#), as well as in [Table 4.7](#) and [Table 4.8](#). We find that in GPT-4 evaluation, GPT-4 tends to assign the lowest scores to its own answers compared to those generated by other fine-tuned models. Summaries written by humans receive the highest scores and are generally regarded as the highest quality. In line with human evaluation findings, GPT-4 also recognizes LoRA as yielding inferior outcomes. In addition, the RST_w^p-LoRA model scored higher than both LoRA and FFT.

Error Analysis. The passage discusses a comparison between RST-LoRA and GPT-4 in generating summaries using two samples from the eLife test set, as shown in [Table 4.9](#) and [Table 4.10](#). It is observed that the summaries generated by GPT-4 tend to be overly general, lacking relevance and informativeness to the source document, and often include the model’s own judgments or inferences. For instance, in the example provided, GPT-4 deviates from the main discussion topic and fails to align the generated summary with factual consistency from the source document. These findings suggest that general-purpose LLMs like GPT-4, without task-specific fine-tuning, still tend to focus more on language fluency but struggle to maintain fidelity to factual information with the source document.

Model	Output
Target-1	<p>Imagine a gymnastics competition in which participants take turns to cartwheel and somersault across the floor. The routines on display comprise sequences of precisely timed movements learned through practice. This is also true for many of the actions we perform every day, such as reaching for a cup of coffee. A region of the brain called the cerebellum helps us learn sequences of movements. But how does it do this? To find out, Khilkevich et al. came up with a new version of an old experiment. Rabbits were first trained to blink their eye in response to a specific external cue. This type of learning, called associative learning, has been shown before in the cerebellum. But Khilkevich et al. wondered whether the cerebellum could also use internal feedback signals from the eye-blink as a cue to learn the next movement. If so, this might explain how the cerebellum can chain movements together in a sequence. As predicted, Khilkevich et al. found that rabbits could learn to blink their eye in response to an initial signal, and then blink again in response to the first blink. Control experiments confirmed that the second eye-blink was coupled to the first, and not to the original cue. Moreover, on many trials, the rabbits showed a third and even fourth eye-blink. This is because feedback signals from the first, second, or third blink were the same. Thus, the feedback signals from the first blink triggered the second blink, feedback from the second triggered the third, and so forth. Rabbits could also learn to use a blink of the left eye as a cue for a blink of the right eye. Similar patterns of neuronal activity accompanied each blink, suggesting that the same mechanism generated them all. The cerebellum can thus use feedback from one movement as a cue to learn the proper timing of the next movement in a sequence. A key question is whether this mechanism of sequence learning extends beyond movement. The cerebellum has extensive connections to the brain's outer layer, the cortex, including many areas involved in cognition. Future experiments should test whether the cerebellum might help guide sequences of cortical activity during cognitive tasks.</p>
GPT-4 _{ICL}	<p>In a recent study, researchers explored the idea of using rabbits to understand human movement, focusing on how the cerebellum, a brain part, might be involved. They looked at how rabbits respond to various light and sound cues, seemingly to understand if these animals could learn complex dance-like moves. The study, which seemed to have a loose structure, suggested that the rabbits' performance in these tasks could somehow be linked to human learning of sequences, like playing a musical instrument or participating in sports. However, the connection between the rabbits' activities and human motor skills was not established, leaving the real impact of the study on human neuroscience quite ambiguous. Additionally, the research didn't delve deeply into the differences between rabbit and human brains, making its applicability to human subjects questionable.</p>
Vicuna _{RST_v^p-LoRA}	<p>The cerebellum's role in movement and learning was explored in a study by Khilkevich et al., where rabbits were trained to blink in response to cues. This research investigated the cerebellum's capacity for associative learning and its ability to use internal feedback from one movement as a cue for the next. Findings revealed that rabbits could learn sequential blinking, with each blink triggering the subsequent one. This suggests that the cerebellum not only coordinates simple movements but can also chain complex sequences of movements together. The study provides insights into the cerebellum's broader role in learning and cognition, particularly in understanding how it processes and coordinates the timing of sequential movements.</p>

Table 4.9: Comparison of target summary with RST-LoRA and GPT-4 outputs.

Model	Output
Target-2	<p>To grow and multiply, a living cell must take a variety of factors into account, such as its own energy levels and the availability of nutrients. A protein called mTOR sits at the core of a signaling pathway that integrates these and other sources of information. Problems with the mTOR pathway contribute to several diseases including diabetes and cancer. The mTOR protein occurs in two distinct protein complexes, called mTORC1 and mTORC2. These complexes contain a mix of other proteins – known as accessory proteins. They also sense different cues and act upon distinct targets in the cell. Recent research reported the structure of mTORC1, which provided clues about how this complex works. Yet, much less was known about the mTORC2 complex. Stutfeld, Aylett et al. have now used a technique called cryo-electron microscopy to reveal the three-dimensional architecture of the human version of mTORC2. Comparing the new mTORC2 structure to the existing one for mTORC1 showed that they have many features in common but important differences too. The overall shape of both complexes is similar and each complex contains two copies of mTOR arranged similarly. Also, the main accessory proteins in each complex interact with almost the same parts of mTOR, but the accessory proteins in mTORC2 are organized differently from those of mTORC1. The different accessory proteins also have distinct shapes. These differences could help to explain why the complexes respond to different cues and recognize different targets. These new findings provide an entry point for further studies on how mTORC2 works in cells. The next step is to get a higher-resolution image of the structure of this complex to see the finer details of all the components. This may in the future help scientists to develop drugs that inhibit mTORC2 to treat cancer and other diseases.</p>
GPT-4 _{ICL}	<p>In this paper, researchers delve into the intriguing world of cellular biology, focusing on a protein known as the mammalian target of rapamycin (mTOR). This protein, often likened to a cellular conductor, orchestrates a symphony of biological processes, guiding cells through complex growth and metabolic pathways. The study particularly highlights the dual roles of mTOR in different cellular complexes, drawing attention to its versatile nature. Interestingly, the paper touches upon the enigmatic behavior of mTOR in various diseases, hinting at its Jekyll and Hyde personality in cellular functioning. This study provides a glimpse into the intricate ballet of molecular interactions, painting a picture of mTOR as a pivotal figure in the dance of life.</p>
Vicuna _{RST_w^p-LoRA}	<p>The cells of animals, yeast, and other eukaryotes all contain a wide range of proteins that carry out different roles. A protein called mTOR sits at the core of this protein, where one letter of the protein sits at one end and the other at the other. The number of mRNAs is so small that it is thought that mTOR is only recently shown to be present in living cells. However, mTOR has only been studied in the laboratory in the form of single-celled structures called human cells. This has made it impractical for researchers to dissect the structure of the mTOR protein and its corresponding protein. Now, Wassagano et al. have used a technique called cryo-electron microscopy to reveal the three-dimensional architecture of mTOR proteins. The experiments reveal that this complex contains a small region called the ‘active’ site that is responsible for producing mTOR. This complex is also involved in the release of energy from blood sugar in cells and is also responsible for many of the processes that occur in the pancreas and immune responses in cells. In particular, the active state of the innermost mTOR complex is activated when the cell is starved for nutrients. Further experiments showed that the active site is involved in both the production and the activity of a protein called sIN1. Wassagani et al.’s findings show that both mTOR and SIN1 have a similar active site to each other in human cancers. These findings shed new light on the role of how mTOR interacts with other proteins in human cells, which may help to identify new targets for developing treatments against these kinds of cancer.</p>

Table 4.10: Comparison of target summary with RST-LoRA and GPT-4 outputs.

4.5 Related Work

Document Summarization with RST. RST is a linguistic discourse framework that provides a way to organize text into a hierarchical tree structure, which helps to better understand the overall organization and inter-part relations of text. Early research by [Marcu \(1997\)](#) and [Louis et al. \(2010\)](#) uncovered that human-written summaries often align with the nucleus EDUs in RST trees. This correlation underscores the validity of RST as a theoretical motivation for summarization tasks. Building on this insight, subsequent studies have demonstrated the value of explicitly incorporating RST trees into neural summarization models. For example, [Kikuchi et al. \(2014\)](#) boosted the summarization performance of the RNN-based model by constructing RST trees, where satellite EDUs were pruned to retain only the nucleus EDUs, thus focusing on the document’s key content. Pre-trained language models have a noted tendency to capture some superficial aspects of discourse relations without explicit training ([Qian et al. 2021](#); [Schuster et al. 2022](#)), but the latent discourse information is often not captured correctly. To alleviate this challenge, [Xu et al. \(2020\)](#) and [Dong et al. \(2021\)](#) enhanced summarization models by incorporating discourse structure within transformer-based and graph neural network models, respectively.

More recently, [Liu et al. \(2023c\)](#) proposed an approach that incorporates the uncertainty of RST structures into the attention mechanisms of summarization models and achieved SOTA results on multiple datasets. However, all the above approaches require full fine-tuning of NLG models, which is very expensive. As the model parameters increase, this issue will be further amplified. Incorporating RST into PEFT might potentially lower the barrier to fine-tuning by structuring the learning process around the inherent rhetorical patterns in the data.

Document Summarization with LoRA. LoRA, presented by [Hu et al. \(2022\)](#), is a low-rank approximation strategy that reduces the number of trainable parameters by freezing the pre-trained model weights and injecting trainable rank decomposition matrices into each layer of the transformer architecture. The initial research also demonstrated through summarization tasks that applying LoRA on the GPT-3 model ([Brown et al. 2020](#)) with less than 1% of the parameters could even outperform FFT. Expanding on this, studies by [Dettmers et al. \(2023\)](#), [Xu et al. \(2024\)](#), and [Li et al. \(2024d\)](#) enhanced generalization ability in downstream summarization tasks by quantifying LoRA matri-

ces and adopting mixed-precision techniques. Furthermore, [Zhu et al. \(2023\)](#) combined LoRA with layer pruning, achieving notable improvements in specialized applications like medical report summarization. Recently, [Liao et al. \(2023\)](#) validated the feasibility of using task-neutral sparse masks to improve the performance in text summarization with LoRA.

In a similar work, [Ghazvininejad et al. \(2022\)](#) integrated hierarchical document structure (i.e., blocking structure) into prefix-tuning to simulate the high-level discourse relation and achieved improvements in the task of text generation. However, there is still an unexplored potential in explicitly integrating fine-grained RST structures into the summarization process with PEFT methods, since comprehending the coherence of discourse elements could positively impact the quality of generated summaries ([Li et al. 2016b](#); [Liu et al. 2019b](#); [Huang et al. 2021](#)), particularly in the context of summarizing long documents ([Cohan et al. 2018](#); [Xu et al. 2020](#); [Li et al. 2020c](#); [Gabriel et al. 2021](#); [Dong et al. 2021](#); [Balachandran et al. 2022](#); [Liu et al. 2023c](#)).

4.6 Chapter Summary

This chapter addressed Research Question 2 by examining whether discourse signals can still exert a measurable influence on content selection and inter-sentential organization when the parameter update budget is severely constrained. The proposed RST-LoRA model injects compact RST distribution matrices into the LoRA training process through element-wise modulation, without increasing the number of trainable parameters. Across four structural variants that vary in granularity (binary vs. probabilistic) and information content (with vs. without relation types), the experiments on Longformer and Vicuna backbones showed that discourse guidance can improve summarization quality even when fewer than 0.5% of model parameters are updated. The variant that retains both relation types and parsing uncertainty consistently performed best, providing empirical support for Hypothesis II and suggesting that structured discourse signals can compensate, to some extent, for the limited expressiveness of low-rank update spaces.

These results extend the findings of the preceding chapter from a full fine-tuning regime to a resource-constrained one, confirming that the distributional representation of discourse structure remains a viable guidance mechanism across different adaptation settings. However, both this chapter and the preceding one operate within a text-to-text summarization interface, where input evidence takes the form of coherent long-form

documents with relatively stable discourse unit boundaries. The next chapter shifts to a setting where evidence is distributed across modalities and where end-to-end generation alone may not be sufficient to reproduce the organizational conventions of the target summary genre, motivating the introduction of explicit planning as an intermediate variable.

5

Discourse-Aware Planning for Video-to-Text Summarization

Chapter 3 and Chapter 4 integrated discourse structure as an actionable supervisory signal into long-document summarization modeling. In particular, Chapter 3 injects type-aware distributed RST representations into sparse attention encoders, while Chapter 4 extends this approach to parameter-efficient fine-tuning settings by softly modulating LoRA’s input representations with discourse coefficient matrices. A shared insight across both chapters is that representing structure as a learnable distributional signal is generally more robust than imposing a single-best parse tree as a deterministic constraint. This insight provides empirical support for Research Question 1 (RQ1), Hypothesis I, Research Question 2 (RQ2), and Hypothesis II. However, these investigations remain grounded in a text-to-text summarization interface, where input evidence consists of coherent long-form documents with stable discourse unit boundaries and reusable rhetorical patterns.

When shifting to the task of summarizing scientific videos, both the form of evidence and the complexity of information organization change substantially. In the multimodal setting, evidence is dispersed across visual scenes, spoken narration, and on-screen text; key arguments unfold across temporally distant segments; and visual elements such as figures and equations substantially affect comprehensibility. These properties pose a challenge for the RST-based distributional approach developed in the preceding chapters. The LDD tensor representation relies on an external RST parser that segments text into elementary discourse units and assigns relation labels with associated confi-

dence scores. In the multimodal setting, however, the input is no longer a well-formed document but a temporally ordered stream of heterogeneous signals, and existing RST parsers are not designed to operate over such inputs. Even if one were to parse a speech transcript, the resulting discourse structure would reflect only the verbal channel and miss information conveyed visually. More fundamentally, the challenge in this setting is not only which local relations hold between adjacent segments, but how to organize an entire summary at the macro level, deciding what rhetorical moves to make, in what order, and with what communicative intent. This is a planning problem rather than a local attention modulation problem, and it calls for a different type of discourse-level intervention.

The Question under Discussion (QUD) framework offers a natural fit for this purpose. Rather than annotating pairwise relations between text spans, QUD conceptualizes discourse progression as a sequence of implicit questions that are raised and resolved over time (Roberts. 2012; Beaver et al. 2017). This perspective aligns well with the structure of scientific abstracts, which typically follow a recognizable sequence of rhetorical moves. By representing the plan as an ordered sequence of questions, each corresponding to a sentence-level intent in the target summary, the model gains an explicit organizational scaffold that can guide generation without requiring a full RST parse of the multimodal input. Against this backdrop, Research Question 3 (RQ3) asks how discourse-aware planning can be extended to support macro-level organization and factual grounding in video-to-text summarization, given the challenges of fragmented and multimodal input.

To this end, Chapter 5 tests Hypothesis III that, when cross-modal alignment is sufficiently reliable, discourse-aware planning enables summarization models to produce more coherent and factually grounded outputs by structuring generation around an abstract representation of argumentative intent. Concretely, Chapter 5 first introduces the VISTA dataset, which consists of 18,599 aligned video–summary pairs where conference presentation videos are paired with their corresponding paper abstracts and spans multiple scientific venues in computational linguistics and machine learning. This resource addresses a gap in existing video summarization research, where high-quality aligned datasets in scientific contexts remain scarce, making it difficult for models to learn how to reorganize multimodal evidence into the conventional organizational logic of scientific abstracts. Building on this foundation, Chapter 5 proposes a planning-based generation framework that draws on the QUD discourse framework. The approach in-

roduces an intermediate planning stage in which a sequence of ordered questions outlines the target summary’s latent structure, and the final summary is then generated conditioned on this plan. Subsequent experiments benchmark both proprietary and open-source multimodal large language models and evaluate whether the plan-based approach can better align with scientific abstract conventions and yield measurable gains in coherence and factual accuracy. [Chapter 5](#) also compares end-to-end generation against planning-based variants and analyzes their relative impact on summary quality and structural organization.

5.1 Problem Formulation

Large multimodal models (LMMs), which integrate components from different modalities through cross-modal alignment training ([Koh et al. 2023](#); [Cheng et al. 2023](#); [Li et al. 2024a](#); [Ahn et al. 2024](#); [Fu et al. 2024](#); [Wu et al. 2024a](#)), have achieved considerable progress in video-to-text summarization tasks for general-purpose content such as YouTube, movies, and news videos ([Li et al. 2020b](#); [Lin et al. 2024b](#); [Krubiński et al. 2023](#); [Hua et al. 2025](#); [Chen et al. 2024a](#); [Zhang et al. 2024a](#); [Qiu et al. 2024](#); [Patil et al. 2024](#); [Mahon et al. 2024a](#); [Mahon et al. 2024b](#)). However, many recent studies have highlighted that these LMMs exhibit reduced performance in scientific contexts, particularly when processing technical terminology and scientific visual elements like figures and tables ([Li et al. 2024b](#); [Lu et al. 2024](#); [Yue et al. 2024](#); [Bai et al. 2024](#); [Liang et al. 2024a](#); [Patil et al. 2024](#); [Huang et al. 2025b](#)). This performance gap might be largely attributed to the absence of specialized datasets for multimodal scientific content ([Chen et al. 2024c](#); [Hu et al. 2024](#); [Pramanick et al. 2024](#); [Zhang et al. 2024b](#)).

Thus, we introduce **VISTA** (**V**ideo to **S**cientific **A**bstract), an English dataset for video-to-text summarization in scientific domains. VISTA consists of 18,599 aligned pairs of conference presentation recordings and their corresponding paper abstracts, collected from leading conferences in computational linguistics ([ACL Anthology](#) including ACL, EMNLP, NAACL, EACL, Findings of *ACL) and machine learning ([ICML](#) and [NeurIPS](#)). [Figure 5.1](#) illustrates an example selected from VISTA. We use the abstract of the paper as a proxy for the summary of the video and benchmark VISTA using several state-of-the-art (SOTA) large models, including closed-source LMMs (Claude 3.5 Sonnet, Gemini 2.0, GPT-o1), as well as open-source LMMs (Video-LLaMA, Video-ChatGPT, mPLUG-Owl3, etc.; [Zhang et al. 2023a](#); [Maaz et al. 2024](#); [Lin et al.](#)

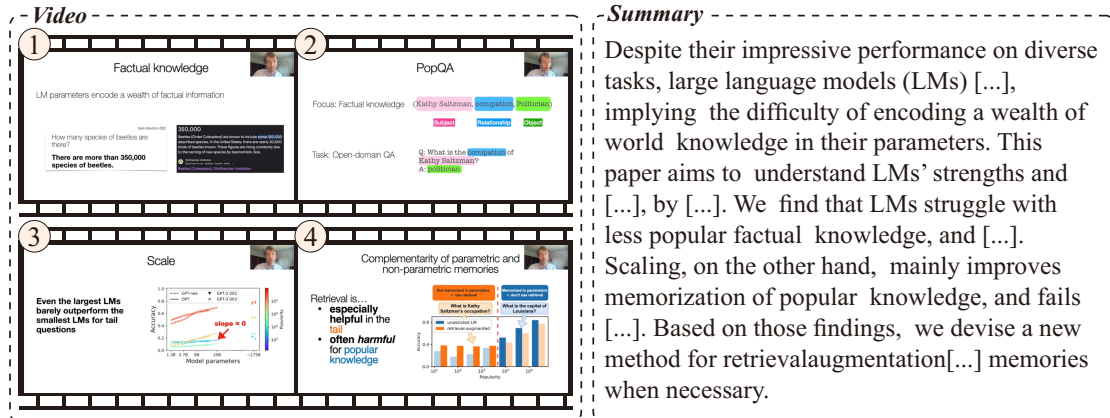


Figure 5.1: An example from VISTA: a conference presentation video (left) paired with the abstract of the corresponding paper (right). This data sample (Mallen et al. 2023) was presented at ACL 2023 and received the Best Video Recordings award.

2024a; Ye et al. 2025; Li et al. 2025a; Li et al. 2024c). For comparison, we also include strong baselines: text-to-text model LLaMA-3.1 (Touvron et al. 2023) and audio-to-text model Qwen2-Audio (Chu et al. 2024). Experiments across zero-shot, QLoRA, and full fine-tuning settings reveal that in-domain fine-tuning improves summarization performance across different large models, and video-based models generally outperform text- and audio-based models on our dataset. However, end-to-end approaches may often struggle to capture the underlying structure of scientific abstracts (Liu et al. 2025b).

To address this, we explore a plan-based approach, which has been shown to improve coherence and factual grounding through a predefined planning component (Narayan et al. 2021; Narayan et al. 2023; Liu et al. 2025b). Unlike direct end-to-end generation, plan-based methods can leverage the fact that scientific abstracts often follow a well-defined format (Takeshita et al. 2024). By explicitly modeling the latent structure of the summary through a sequence of intermediate plans, the summary generation process can be better guided. Empirical results confirm that the plan-based method outperforms existing SOTA models in terms of summary quality and factual accuracy. This work also lays the groundwork for future investigations into the multimodal summarization of scientific videos.

5.2 The VISTA Dataset

Data Acquisition and Cleaning. VISTA is derived from computational linguistics and machine learning conferences, including [ACL Anthology](#) (ACL, EMNLP, NAACL, EACL, Findings of *ACL), [ICML](#), and [NeurIPS](#), covering content from 2020 to 2024. All materials (paper abstracts and video recordings) are contributed by the respective paper authors, ensuring narrative consistency. Since these metadata are stored in XML/JSON files on their respective websites, no further data preprocessing (e.g., extracting abstracts from PDFs) is required. We collect paper titles, author lists, paper abstracts, links to papers, and presentation videos, in accordance with platform terms for academic research purposes (or obtain written confirmation). To maintain one-to-one video-to-text alignments, we exclude samples that may cover multiple papers (e.g., tutorials, invited talks) and videos shorter than one minute or longer than 30 minutes.

Copyright Discussion. According to the statement displayed on the [ACL Anthology](#) website, “*Permission is granted to make copies for the purposes of teaching and research*”, allowing us to use the corresponding data. For [ICML](#) and [NeurIPS](#), we (the authors) have obtained written confirmation granting permission to use the paper titles, author lists, paper abstracts, full papers, and presentation videos available on their websites for research purposes.

Quality Control. We verify the data quality through both manual and automated checks.

- **Manual Control:** We randomly select 500 video-summary pairs to assess whether the summaries provide accurate descriptions of the videos. Two Ph.D. candidates in Computer Science or Computational Linguistics perform binary judgments on these pairs. Across all 500 samples, neither evaluator rejected any sample.
- **Automated Control:** To go beyond the limited scope of manual checks, we employ GPT-o1 for automated assessment using the same binary criteria across all data samples. The model initially flagged 39 pairs as potentially invalid. These flags were likely caused by difficulties in interpreting domain-specific terms or rare expressions and sensitivity to variations in summary length. After further manual review, all 39 samples were confirmed as valid and retained in the dataset.

Data Splits. After quality control, our dataset comprises 18,599 samples, with venue distributions shown in [Figure 5.2](#). To ensure balanced domain coverage in each subset, we proportionally sample to split the dataset into training (80%), validation (10%), and test (10%) sets. All subsequent experiments are conducted using these splits.

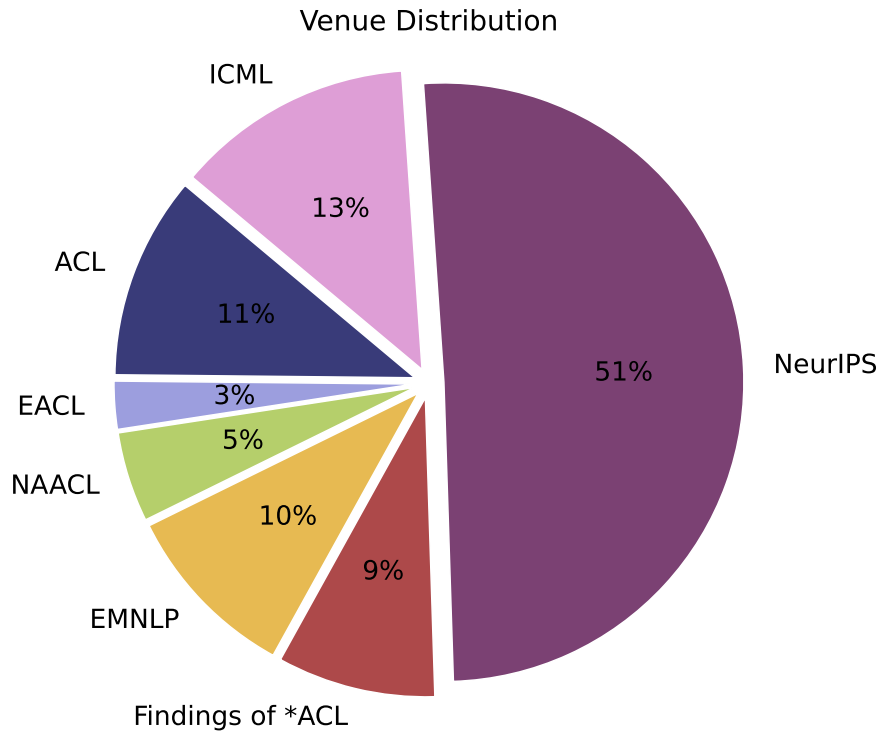


Figure 5.2: Venue distribution of the VISTA dataset.

Dataset Comparison and Statistics. [Table 5.1](#) compares VISTA with several existing video-to-text summarization datasets. While many focus on open-domain (e.g., MM-Sum, Instruct-V2Xum) or areas like news (MLASK, MM-AVS) and activities (VideoXum), VISTA is tailored for summarizing scientific presentations. On average, it features longer inputs (6.8 minutes) than VideoXum (2.1 minutes) and MSS (3.4 minutes), as well as longer summaries (192.6 tokens), compared to YouCook2 (67.8 tokens) and VideoXum (49.9 tokens).

[Table 5.2](#) summarizes the VISTA dataset statistics: Videos average 6.76 minutes and 16.36 shots (we use `PySceneDetect` with `ContentDetector` to calculate video shots), while summaries contain 192.62 tokens on average across 7.19 sentences. The aver-

Dataset	Language	Domain	#Videos	VideoLen	SumLen
MSS (Li et al. 2017)	English, Chinese	News	50	3.4	—
YouCook2 (Zhou et al. 2018)	English	Cooking	2.0K	5.3	67.8
VideoStorytelling (Li et al. 2020a)	English	Open	105	12.6	162.6
VMSMO (Li et al. 2020b)	Chinese	Social Media	184.9K	1.0	11.2
MM-AVS (Fu et al. 2021)	English	News	2.2K	1.8	56.8
MLASK (Krubiński et al. 2023)	Czech	News	41.2K	1.4	33.4
VideoXum (Lin et al. 2024b)	English	Activities	14.0K	2.1	49.9
Shot2Story20K (Han et al. 2025b)	English	Open	20.0K	0.3	201.8
BLiSS (He et al. 2023)	English	Livestream	13.3K	5.0	49.0
SummScreen ^{3D} (Papalampidi et al. 2023)	English	Open	4.5K	40.0	290.0
Ego4D-HCap (Islam et al. 2024a)	English	Open	8.3K	28.5	25.6
Instruct-V2Xum (Hua et al. 2025)	English	Open	30.0K	3.1	239.0
MMSum (Qiu et al. 2024)	English	Open	5.1K	14.5	21.7
LfVS-T (Argaw et al. 2024)	English	YouTube	1.2K	12.2	—
VISTA (ours)	English	Academic	18.6K	6.8	192.6

Table 5.1: Comprison of video-to-text summarization datasets. #Videos = the number of videos, whereas VideoLen and SumLen refer to the average of video duration (in minutes) and the average number of summary tokens.

age dependency tree depth (Avg. Depth of Dep Tree) is 6.02, indicating the syntactic complexity of the summaries. Meanwhile, the Type-Token Ratio (TTR) is 0.62, reflecting lexical diversity. Both metrics are calculated using *spaCy*. Diversity metrics (Li et al. 2016a), which measure the variety of unique n-grams, yield Distinct-1, Distinct-2, and Distinct-3 scores of 0.62, 0.93, and 0.97, respectively. Figure 5.3 visualizes key attributes: Most summaries remain under 250 tokens and 10 sentences, and most videos last fewer than 10 minutes with under 30 shots. In Appendix Section C.1, we present a random sample from the VISTA dataset.

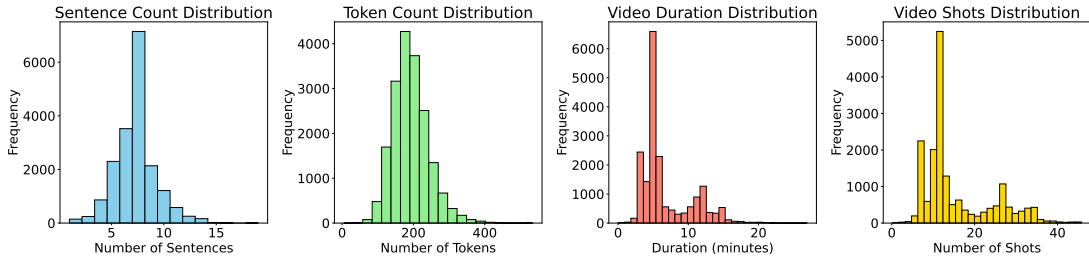


Figure 5.3: Distribution of summary sentences, summary tokens, video durations, and video shots in VISTA.

Training / Validation / Test Set	14,881 / 1,859 / 1,859
Avg. Video Length (mins) / Shots	6.76 / 16.36
Avg. #Summary Sent / Tokens	7.19 / 192.62
Avg. Depth of Dep Tree	6.02
Type-Token Ratio	0.62
Distinct-1 / -2 / -3	0.62 / 0.93 / 0.97

Table 5.2: Key statistics of the VISTA dataset, showcasing the average video length and shot count, summary characteristics (sentence and token counts), syntactic complexity (dependency tree depth), and lexical diversity (Type-Token Ratio and Distinct n-gram scores).

5.3 Benchmarking VISTA

Task Overview. We formalize the task of summarizing recorded scientific videos as follows: Let v and s denote a video (or its transcript/audio) and its paired summary from dataset $D = \{(v_1, s_1), (v_2, s_2), \dots, (v_n, s_n)\}$, where n signifies the number of video-summary pairs. The objective is to train a (multimodal) model \mathcal{M} to learn the conditional probability distribution $P(s | v)$. Given a new video, the trained model \mathcal{M} is expected to generate an appropriate summary.

A challenge in video-to-text summarization is structuring the generated summaries in a coherent and faithful manner. Directly learning the mapping from v to s could lead to inadequate outputs, as the model lacks explicit guidance on how to organize and present the extracted information (Mahon et al. 2024a). Scientific abstracts often follow a relatively well-defined structure, making them suitable for a more structured generation approach (Takeshita et al. 2024). We follow previous work (Narayan et al. 2021; Narayan et al. 2023) in adopting a plan-based framework that introduces an intermediate representation to capture latent structure more effectively than simpler end-to-end approaches. Specifically, given input v , we first generate a plan p , which consists of a sequence of automatically generated questions $\{q_1, q_2, \dots, q_m\}$, each corresponding to a sentence to be verbalized in the summary. The plan explicitly controls the structure of the summary as a whole and the content of each of its sentences (which are meant to answer the questions in the plan). The model is then trained to learn the extended conditional probability distribution $P(s | v, p)$, ensuring that the generated summaries follow the structure and flow of plan p .

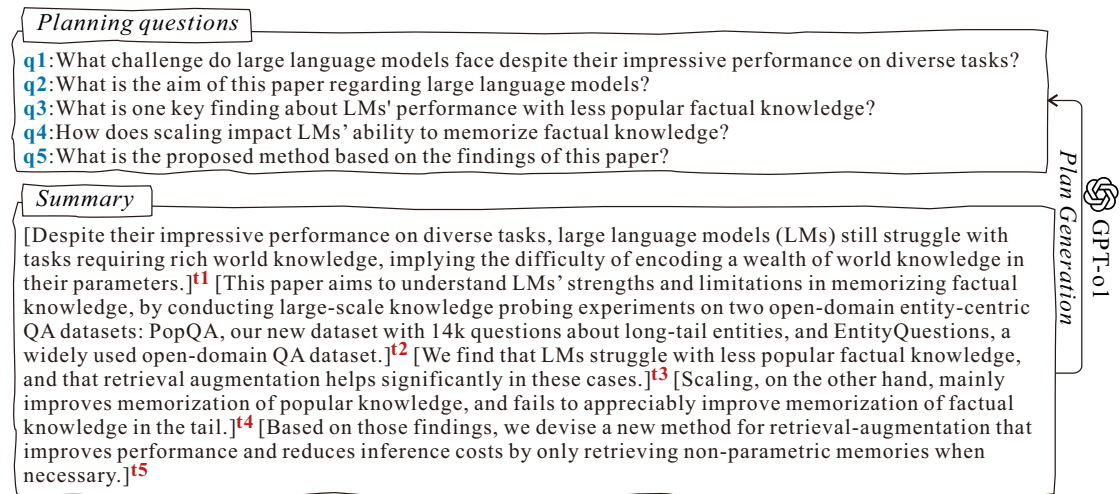


Figure 5.4: GPT-o1 generates plans based on reference summaries. Each question q_i corresponds to a summary sentence t_i , which we assume constitutes its answer. Index i ranges from 1 to the number of summary sentences.

Plan Generation. We hypothesize that summary sentences can be viewed as responses to plan questions, where the plan consists of an ordered sequence of questions directly associated with the target content. This idea is inspired by the theory of Question Under Discussion (QUD; Roberts (2012), Wu et al. (2023b), and Suvarna et al. (2024)), which posits that discourse often revolves around a set of questions that guide the structure and interpretation of the conversation.

We leverage GPT-o1 (OpenAI. 2023) to generate silver-standard plans based on reference summary sentences and their preceding context. As shown in Figure 5.4, for example, question q_3 is generated based on target sentence t_3 and the summary sentences preceding it (i.e., t_1 and t_2), and so on. As a result, the question sequence preserves the order of sentences in the reference summaries, ensuring that the plan maintains a natural and coherent flow consistent with the structure of reference summaries. The prompt used to generate plan questions is provided in Appendix Figure C.4.

Summarization Model. We train two independent modules corresponding to Plan Generation (PG) and Summary Generation (SG). The PG module is trained on pairs of (v, p) samples. The SG module is trained on tuples $([v; p], s)$, where $[v; p]$ is the concatenation of the input v and its plan p . During inference, the trained PG module predicts plan \hat{p} for input v , and the tuple $[v; \hat{p}]$ is fed into the SG module to generate the final summary. Both modules have the same backbone but are trained independently.

5.4 Experimental Setup

Baseline Models. We benchmark our dataset using three learning settings: Zero-shot learning, QLoRA fine-tuning (Dettmers et al. 2023), and full-parameter fine-tuning. For zero-shot learning, we test closed-source multimodal models, including GPT-o1 (OpenAI. 2023), Gemini 2.0 (Team. 2023), Claude 3.5 Sonnet (Anthropic. 2024), as well as open-source LMMs such as Video-LLaMA (Zhang et al. 2023a), Video-ChatGPT (Maaz et al. 2024), Video-LLaVA (Lin et al. 2024a), LLaMA-VID (Li et al. 2024c), LLaVA-NeXT-Interleave (Li et al. 2025a), and mPLUG-Owl3 (Ye et al. 2025). These open-source video LMMs process videos by extracting multimodal features, such as visual and/or audio components, using cross-modal attention mechanisms to align and integrate information across modalities.

We also assess LLaMA-3.1 (Touvron et al. 2023) and Qwen2-Audio (Chu et al. 2024) to examine if text- or audio-based models can accomplish the summarization task without taking video information into account. For LLaMA-3.1, we explore two variants: In LLaMA-3.1_{transcript}, we extract audio from video files using `moviepy` and transcribe it with OpenAI’s `Whisper-1` to generate text input for the model. In LLaMA-3.1_{OCR}, we apply `EasyOCR` to extract on-screen text from video frames and use the OCR-generated text as input for summarization. Similarly, for Qwen2-Audio, we use `moviepy` to convert video files into audio and treat the audio as input. Exact model versions are provided in Table 5.3. Based on our benchmarking results, we select the best-performing model as the backbone for the plan-based strategy and evaluate its performance. Prompts for the above models are offered in Appendix Section C.2 (Figures C.3–C.6).

Hyper-parameters Settings. For all fine-tuning experiments, we utilize the AdamW optimizer (Loshchilov et al. 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$, and a weight decay of 0.1, combined with a warm-up ratio of 0.15. The initial learning rate is set to $5e-5$, with cosine learning rate scheduling. DeepSpeed is configured with ZeRO-3 Offload. We set the random seed to 2025 and apply a dropout rate of 0.1. In the QLoRA setting, the rank r is set to 32, the scaling factor α is set to 64, and the dropout rate for the low-rank matrices is 0.1. All other parameters follow the default settings of the Transformers library.

Model	Version	Model Size
GPT-o1 (OpenAI. 2023)	o1-2024-12-17	Unknown
Gemini 2.0 (Team. 2023)	Gemini 2.0 Flash	Unknown
Claude 3.5 Sonnet (Anthropic. 2024)	claude-3-5-sonnet-20241022	Unknown
LLaMA-3.1 (Touvron et al. 2023)	LLaMA-3.1-8B-Instruct	8B
Qwen2-Audio (Chu et al. 2024)	Qwen2-Audio-7B-Instruct	7B
Video-LLaMA (Zhang et al. 2023a)	VideoLLaMA2-7B-16F	7B
Video-ChatGPT (Maaz et al. 2024)	Video-ChatGPT-7B	7B
Video-LLaVA (Lin et al. 2024a)	Video-LLaVA-7B-hf	7B
LLaMA-VID (Li et al. 2024c)	LLaMA-VID-7B-Full-224-Long-Video	7B
LLaVA-NeXT-Interleave (Li et al. 2025a)	LLaVA-NeXT-Interleave-Qwen-7B	7B
mPLUG-Owl3 (Ye et al. 2025)	mPLUG-Owl3-7B-241101	7B

Table 5.3: Model version details.

During training, we save the checkpoint with the highest Rouge-2 F1 score on the validation set as the final model. All experiments are conducted over 16 epochs with a batch size of 16 and early stopping (all models converged before 16 epochs). For model inference (including zero-shot learning), we employ a beam search with a beam of size 4, a length penalty of 3.0, a no-repeat n-gram size of 3, and the maximum number of new tokens generated is limited to 256. For video-based LMMs, the sampling rate is set to 0.1 fps, and the number of extracted frames is set to 32.

For closed-source models, results are obtained via API requests during the experimental period from 01/09/2024 to 10/02/2025. The hyperparameter settings for these API requests include a temperature of 1, top_p of 1, a frequency penalty of 0.2, and a presence penalty of 0.2. All other parameters adhere to the default settings specified by their respective platforms. To ensure a fair comparison, all models, including baselines, plan-based models, and ablation models, are tested using identical prompt instructions.

Evaluation Metrics. We report a set of evaluation metrics to measure informativeness, alignment, and factual consistency in summaries. For informativeness, we utilize ROUGE (Lin. 2004), SacreBLEU (Post. 2018), METEOR (Banerjee et al. 2005), BERTScore (Zhang et al. 2020b), and CIDEr-D (Vedantam et al. 2015). Specifically, we provide the F1 scores for Rouge-1 (R1), Rouge-2 (R2), and Rouge-LSum (RLSUM). Alignment to the input video is evaluated with VideoScore (He et al. 2024), and factual consistency with FactVC (Liu et al. 2023d).

5.5 Results and Analysis

Plans Quality Validation. To validate the quality of the silver-standard plans generated by GPT-o1, we conduct a manual evaluation on 100 randomly selected samples. The evaluation is carried out by the same annotators involved in our human evaluation setup. Each annotator is asked to make a binary judgment on whether the generated plan question satisfied two validity criteria: (1) Local Coherence: The question is well-formed and semantically related to the summary; and (2) QUD-Alignment: Each sentence in the summary could plausibly serve as an answer to the question, consistent with the QUD framework. We observe strong inter-annotator agreement (Fleiss’ $\kappa = 0.853$), indicating a high degree of consistency in decisions. In addition to this, we perform a manual error analysis to screen for systematic biases or recurrent flaws, such as overly generic phrasing, hallucinated entities, or structural redundancy. No such patterns are observed.

General Results. Table 5.4 compares model performance across three learning settings: Zero-shot, QLoRA fine-tuning, and full-parameter fine-tuning. Overall, fine-tuning on in-domain data yields substantial performance gains across all evaluation metrics. Full fine-tuning consistently outperforms QLoRA. While closed-source models such as GPT-o1 and Gemini typically lead in zero-shot performance, open-source models like mPLUG-0w13 and Plan-mPlug-0w13 achieve competitive or even superior results when fine-tuned, especially in semantic alignment (BERTScore) and video-text consistency (VideoScore).

We also observe that video-based LMMs consistently outperform text-based and audio-based models. While models such as LLaMA-3.1_{transcript}, LLaMA-3.1_{OCR}, and Qwen2-Audio yield comparable results, they lag behind video-grounded models in overall performance. In particular, mPLUG-0w13 achieves SOTA results across most metrics, highlighting the crucial role of visual information in enhancing summarization quality.

Plan-mPlug-0w13 is the plan-based approach built on mPLUG-0w13, outperforming all open-source baselines in both zero-shot and fine-tuned settings. For zero-shot inference, the Plan-mPlug-0w13* variant, which fine-tunes only the Plan Generation (PG) module, surpasses other models in summary quality, factual consistency, and semantic alignment. With full-parameter fine-tuning, Plan-mPlug-0w13 achieves the highest overall scores across models, showing improvements in factual accuracy (+3.47 in

Method	Model	Open-source	R1	R2	RLsum	SacreBLEU	Meteor	BERTscore	CIDEr-D	VideoScore	FactVC
Zero-shot Learning	LLaMA-3.1 _{transcript}	✓	23.68	4.22	21.39	2.70	14.62	80.93	1.17	1.53	34.32
	LLaMA-3.1 _{OCR}	✓	24.02	4.37	21.42	2.63	14.59	80.33	1.19	1.50	34.06
	Qwen2-Audio	✓	23.52	4.29	21.53	2.49	14.77	80.62	1.15	1.59	34.31
	Claude 3.5 Sonnet	✗	27.71	5.59	24.14	3.14	17.53	82.57	1.32	1.91	50.11
	Gemini 2.0	✗	27.82	5.66	24.29	4.22	17.83	82.64	1.47	2.02	52.02
	GPT-o1	✗	27.90	5.69	24.37	4.38	17.90	82.63	1.61	2.17	51.36
	Video-LLaMA	✓	20.18	3.19	21.24	1.76	13.73	81.31	1.08	1.63	32.25
	Video-ChatGPT	✓	20.36	3.52	21.43	1.79	14.01	81.35	1.11	1.63	33.21
	Video-LLaVA	✓	25.29	4.50	22.52	2.82	15.13	81.39	1.17	1.65	36.45
	LLaMA-VID	✓	25.31	4.77	22.53	2.88	15.27	81.32	1.14	1.64	36.39
	LLaVA-NeXT-Interleave	✓	25.41	4.82	22.68	2.92	15.25	81.40	1.18	1.73	40.12
	mPLUG-0w13	✓	25.57	4.82	22.84	2.99	15.33	81.39	1.21	1.77	42.07
Plan-mPlug-0w13*	✓	25.62 [†]	4.95 ^{†‡}	22.97 ^{†‡}	3.14 ^{†‡}	15.39 ^{†‡}	81.45 [†]	1.27 ^{†‡}	1.86 ^{†‡}	47.37 ^{†‡}	
QLoRA Fine-tuning	LLaMA-3.1 _{transcript}	✓	32.24	11.38	30.39	8.03	21.57	82.39	3.86	2.81	53.22
	LLaMA-3.1 _{OCR}	✓	33.01	12.11	30.52	8.04	21.55	82.41	3.92	2.77	53.19
	Qwen2-Audio	✓	32.17	12.05	30.77	7.87	21.86	82.36	4.11	2.80	54.27
	Video-LLaMA	✓	30.74	9.44	28.33	6.45	22.49	82.61	3.99	2.77	52.05
	Video-ChatGPT	✓	31.68	10.50	30.40	7.63	23.67	82.62	4.02	2.78	55.02
	Video-LLaVA	✓	33.16	12.64	30.37	8.17	23.92	82.81	4.26	2.83	59.13
	LLaMA-VID	✓	33.31	12.73	30.49	8.22	23.90	83.01	4.31	2.88	62.20
	LLaVA-NeXT-Interleave	✓	33.37	12.77	30.56	8.30	23.95	83.47	4.47	2.93	66.14
	mPLUG-0w13	✓	33.40	12.82	30.66	8.29	23.97	83.49	4.47	2.92	70.08
	Plan-mPlug-0w13	✓	33.52 ^{†‡}	13.01 ^{†‡}	31.10 ^{†‡}	8.33	24.11 ^{†‡}	83.53 [†]	4.52	3.11 ^{†‡}	73.11 ^{†‡}
Full Fine-tuning	LLaMA-3.1 _{transcript}	✓	33.37	11.93	30.86	8.27	25.12	83.71	4.87	3.21	63.38
	LLaMA-3.1 _{OCR}	✓	34.02	12.42	31.72	8.51	15.11	84.09	4.89	3.32	65.84
	Qwen2-Audio	✓	33.82	12.37	31.63	8.33	25.09	83.62	4.83	3.22	66.62
	Video-LLaMA	✓	32.19	11.86	31.68	8.41	24.99	83.83	4.77	3.04	64.21
	Video-ChatGPT	✓	32.47	12.11	32.21	8.72	25.09	83.91	4.82	3.11	66.09
	Video-LLaVA	✓	33.28	13.39	32.78	9.10	25.42	83.97	4.87	3.13	66.12
	LLaMA-VID	✓	33.47	13.53	32.80	9.21	25.41	84.03	4.91	3.17	68.30
	LLaVA-NeXT-Interleave	✓	33.75	13.61	32.88	9.26	25.63	84.11	5.01	3.23	73.42
	mPLUG-0w13	✓	34.22	13.62	32.91	9.32	25.72	84.22	5.03	3.28	71.94
	Plan-mPlug-0w13	✓	34.53 ^{†‡}	13.74 ^{†‡}	33.25 ^{†‡}	9.56 ^{†‡}	25.88 ^{†‡}	84.37 ^{†‡}	5.15 ^{†‡}	3.33 ^{†‡}	75.41 ^{†‡}

Table 5.4: Model performance on VISTA dataset. In Plan-mPlug-0w13*, only the PG module is trained. Plans generated by the PG on the test set serve as input to the SG module for zero-shot inference (no training is applied to the SG module). Symbols [†] and [‡] indicate that the performance of Plan-mPlug-0w13 is significantly ($p < 0.05$) different from LLaVA-NeXT-Interleave (third best) and mPLUG-0w13 (second best), when using the paired t-test.

FactVC) and quality (+0.34 in RLsum) compared to mPLUG-0w13. However, all models (including the plan-based method) exhibit hallucinations (FactVC) and alignment (VideoScore) issues, and there are still significant differences (p-value of the paired t-test is less than 0.05) between the human performance in this task, with reference summaries scoring 88.54 on FactVC and 4.62 on VideoScore.

Impact of Modality Interplay. To explore the impact of different modality combinations on our multimodal tasks, we conduct an experiment using Video-LLaMA (Zhang et al. 2023a). Seven modality combinations are considered, including unimodal inputs (video, audio, transcript) and their pairwise or joint combinations. For each configuration, only the corresponding modality modules are updated while the remaining ones are

Modality	Zero-shot Learning				QLoRA Fine-tuning				Full Fine-tuning			
	R2	RLsum	VideoScore	FactVC	R2	RLsum	VideoScore	FactVC	R2	RLsum	VideoScore	FactVC
Video only	2.68	20.34	1.55	28.93	8.83	27.51	2.65	50.66	10.78	30.02	2.91	60.87
Audio only	2.14	19.72	1.41	26.84	7.52	26.34	2.48	45.79	9.23	27.93	2.73	58.02
Transcript only	2.02	18.01	1.34	25.53	6.91	24.33	2.39	44.87	8.44	25.81	2.35	54.11
Video + Audio	3.19	21.24	1.63	32.25	9.44	28.33	2.77	52.05	11.86	31.68	3.04	64.21
Video + Transcript	1.87	18.94	1.39	27.76	7.35	24.82	2.51	48.63	9.01	27.19	2.65	58.91
Audio + Transcript	1.64	18.55	1.35	27.48	7.23	24.73	2.38	47.15	8.57	25.82	2.54	55.39
Video + Audio + Transcript	1.92	19.13	1.47	28.60	7.37	25.29	2.52	50.72	9.22	27.21	2.61	59.30

Table 5.5: Performance comparison of different modality combinations.

kept frozen. The summarized results are shown in [Table 5.5](#).

The results consistently show that video is the strongest standalone modality, likely due to its rich spatial-temporal information. Audio offers complementary prosodic and timing cues, but lacks semantic visual grounding. The transcript, while semantically rich, often introduces long, noisy, and unstructured inputs, particularly from ASR systems, that can overwhelm the model’s attention and interfere with alignment. These findings suggest that current video-based LMMs face challenges in effectively aligning and fusing token-heavy, noisy textual inputs with corresponding visual or audio information.

Impact of Plan Generation Ablations. We analyze the plan generation ablation by comparing it with simpler baselines: Lead-3_Q, Tail-3_Q, and Random-3_Q. In these ablation baselines, plans are generated by selecting the first three, last three, or three randomly chosen summary sentences, respectively. Each selected sentence serves as a target for generating a question, with its preceding sentences providing the context. For instance, in the Lead-3_Q setting, the first sentence is used as the target (without any preceding context), prompting the first question in the plan, while subsequent sentences incorporate earlier ones as context. Additionally, we compare the case where QUD is not considered. That is, we directly let GPT-o1 generate all plan questions at once based on the reference summary (NoQUD).

[Table 5.6](#) underlines the performance differences across different plan generation ablations. For NoQUD, it underperforms compared to the QUD-based approach. The Lead-3_Q strategy performs better overall compared to Tail-3_Q and Random-3_Q, indicating that initial sentences offer stronger contextual continuity for generating plan questions.

Impact of Plan Quality. We assess how the quality of the plan questions affects model performance. We apply GPT-o1 as a question generator in a zero-shot setting in our pre-

Model	R2	RLsum	VideoScore	FactVC
Plan-mPlug-Owl3	13.74	33.25	3.33	75.41
NoQUD	13.66	33.02	3.28	73.32
Lead-3 _Q	12.87	30.64	2.95	71.26
Tail-3 _Q	11.62	30.51	2.88	63.82
Random-3 _Q	11.57	30.48	2.87	64.28

Table 5.6: Performance comparison of different plan generation ablations under full fine-tuning settings.

vious experiments. For comparative analysis, we additionally incorporate Llama-3.1 and a state-of-the-art question generation algorithm (RAST) from Gou et al. (2023) to generate the plan questions. In addition, we apply a Random Replacement (RR) method, where questions generated by GPT-o1 are randomly replaced with irrelevant ones. The number of replaced questions per summary ranges from one to the entire set. We also introduce full random replacement (FRR), where questions generated by GPT-o1 are all replaced with random irrelevant questions.¹

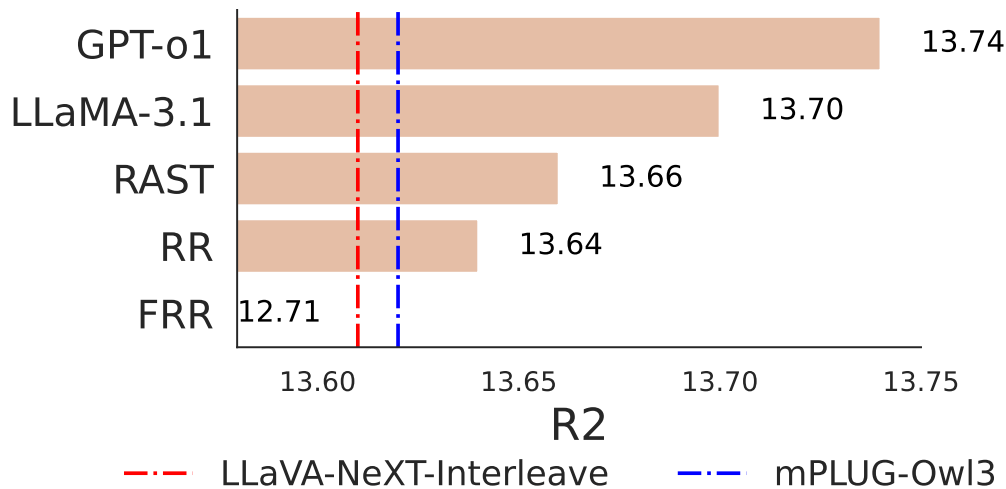


Figure 5.5: Noise in plan generation impacts summarization performance. FRR is a shorthand for Full Random Replacement, and RR for Random Replacement. RAST is a SOTA question generation method.

Figure 5.5 reveals that the quality of plan questions does influence the summarization performance: Using GPT-o1 to generate questions outperforms the rest. The FRR method performs the worst, as irrelevant questions disrupt the alignment between the

¹ The prompt for generating irrelevant questions is given in Appendix Figure C.7.

plan and summary content. We also find that the plan-based method exhibits a certain degree of robustness, as it performs reasonably well even when the plans contain some degree of noise (RR vs. FRR). These findings emphasize the importance of question relevance and quality in structuring the output summaries.

Impact of Video Context on Summary Generation. We examine the impact of different video context configurations on summary generation, comparing mPLUG-0w13 with Plan-mPlug-0w13. Unlike earlier experiments that use the full video as input, here only the first or last 10% or 30% of the video is provided as input. We report results in the full fine-tuning setting.

Context	Model	R2	RLsum	VideoScore	FactVC
All	mPLUG-0w13	13.62	32.91	3.28	71.94
	Plan-mPlug-0w13	13.74	33.25	3.33	75.41
First 10%	mPLUG-0w13	6.31	25.44	2.37	51.02
	Plan-mPlug-0w13	7.37	27.38	2.52	52.39
First 30%	mPLUG-0w13	9.42	28.88	2.78	54.10
	Plan-mPlug-0w13	10.59	30.13	2.78	55.37
Last 10%	mPLUG-0w13	6.53	27.34	2.51	53.64
	Plan-mPlug-0w13	7.62	29.73	2.77	55.93
Last 30%	mPLUG-0w13	7.32	29.17	2.82	57.36
	Plan-mPlug-0w13	10.72	31.29	2.98	62.05

Table 5.7: Model performance under different video context configurations (full fine-tuning). The video content at the end is more helpful for summary generation.

The results in [Table 5.7](#) indicate that partial video context consistently underperforms compared to using the full video. Using the last part of the video generally produces better results than using the first part, as concluding sections often summarize key findings while opening sections primarily introduce background information. Additionally, utilizing 30% of the video outperforms using only 10%, highlighting that more content generally yields better outputs. Across all configurations, the Plan-mPlug-0w13 model consistently outperforms mPLUG-0w13.

Impact of Text Context on Plan Generation. The generation of plan questions in our experiments is influenced by the target sentence and its context. In our main experiments, plan questions are generated based on the target sentence and its preceding summary text (Previous-Context), in line with the original Question Under Discussion

(QUD) requirements (Wu et al. 2023a; Wu et al. 2023b; Liu et al. 2025b). We now assess configurations that generate questions only based on the target sentence (No-Context) or the entire summary (All-Context).

As shown in Figure 5.6, performance differences between different context configurations are relatively small (yet superior to models without planning components shown as red and blue dashed lines). No-Context shows the lowest performance but is the most cost-effective, as it requires the shortest input length for GPT-o1 during question generation. All-Context achieves slightly better results but at the highest computational cost due to the long input length. Previous-Context is aligned with QUD and strikes a good balance, achieving the best performance for a moderate cost.

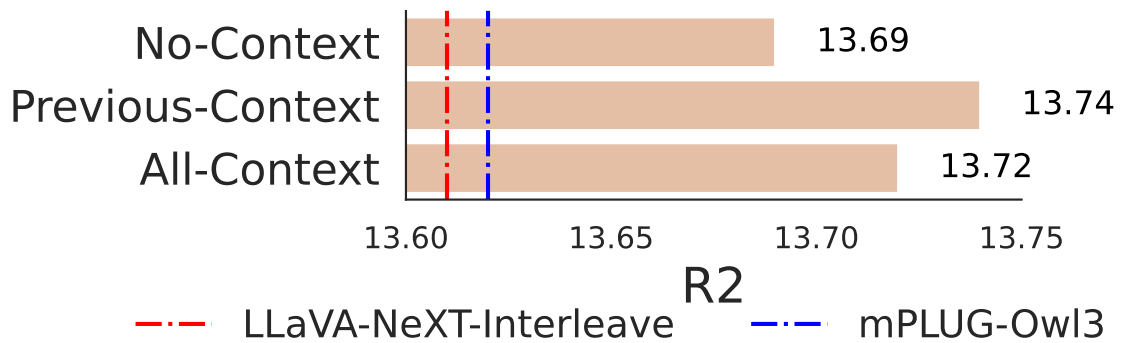


Figure 5.6: Impact of text context for plan generation.

Controllable Generation. An advantage of plan-based models is their ability to control the output summaries by modifying the plans used for generation. We investigate how modifying the structure and composition of these plans impacts the generated summaries, specifically comparing their performance against direct summary generation control through instructions. To this end, we design two controlled experiments:

- *Summary Readability*: How question complexity affects readability, tailored for lay readers or expert readers.
- *Summary Length*: How the number of questions influences summary length, by removing 10%, 30%, and 60% of questions.

We note that the plan-based method employs an explicit planning component where each sentence is guided by a corresponding question that facilitates fine-grained control over the summary’s style or content. Specifically, after PG produces the plan, we use GPT-o1 to edit it and then feed the edited questions back to SG for the final output.

For GPT-o1, which operates in a zero-shot manner, we prepend constraints directly in the prompt. Specifically, GPT-o1 generates an initial summary in one pass and then applies additional prompt-based instructions during a secondary rewriting step to control the output. Both control experiments (Table 5.8 and Table 5.9) reveal similar trends: While performance declines for both models, the plan-based method is more robust and controllable.

Condition	Plan-mPlug-0w13		GPT-o1	
	R2	FRE	R2	FRE
No change	13.74	30.62	5.69	26.37
Lay questions	13.38	35.17	4.26	28.94
Expert questions	13.24	23.54	4.13	24.33

Table 5.8: Control experiment for summary readability. FRE = Flesch Reading Ease.

Condition	Plan-mPlug-0w13		GPT-o1	
	R2	Avg. #Tokens	R2	Avg. #Tokens
No deletion	13.74	202.39	5.69	267.32
Delete 10%	11.05	178.47	4.32	220.49
Delete 30%	10.41	137.72	3.17	192.42
Delete 60%	8.01	100.32	2.98	185.28

Table 5.9: Control experiment for summary length.

In the readability control experiment (Table 5.8), both models show reductions in R2, but Plan-mPlug-0w13 declines less, averaging an R2 loss of 0.43 compared to 1.50 for GPT-o1. Furthermore, Plan-mPlug-0w13 controls readability more effectively, achieving a higher Flesch Reading Ease (FRE) score² of 35.17 for lay questions, compared to 28.94 for GPT-o1, and a lower FRE score of 23.54 for expert questions. In the length control experiment (Table 5.9), R2 scores decline as content is removed, but the plan-based model aligns more closely with target compression ratios, producing summaries averaging 100.32 tokens under 60% deletion, while GPT-o1 generates longer summaries (185.28 tokens).

² The FRE score, which ranges from 0 to 100, measures text readability, with higher scores indicating easier-to-read content, and lower scores reflecting greater complexity.

Planning Beyond Vision. While our primary objective is to evaluate the planning framework in the context of video-to-text summarization, it is valuable to assess its applicability to unimodal, non-visual models. To this end, we conduct supplementary experiments applying the planning method to three models that do not utilize video inputs: (1) LLaMA-3.1_{transcript} (ASR-based textual input), (2) LLaMA-3.1_{OCR} (OCR-based textual input), and (3) Qwen2-Audio (audio-based input). For each model, we compare baseline performance (i.e., without planning) against the planning counterpart. As summarized in Table 5.10, planning consistently improves performance across all settings and evaluation metrics. A paired t-test confirms that these improvements are statistically significant ($p < 0.05$).

These findings demonstrate that the planning method does not function solely as a domain-specific enhancement but rather as a generalizable scaffold that supports better discourse structure, even in the absence of visual input. We hypothesize that, for text- and audio-based models, planning mitigates the lack of spatial-temporal signals by providing discourse-level anchors, such as intent-driven prompts (e.g., “What problem is being addressed?”), that guide the model’s summarization trajectory.

Model	Setting	R2	RLsum	VideoScore	FactVC
LLaMA-3.1 _{transcript}	Zero-shot Learning	4.22 → 4.56	21.39 → 22.01	1.53 → 1.75	34.32 → 40.78
	QLoRA Fine-tuning	11.38 → 11.62	30.39 → 30.55	2.81 → 3.02	53.22 → 60.47
	Full Fine-tuning	11.93 → 12.24	30.86 → 31.38	3.21 → 3.25	63.38 → 65.21
LLaMA-3.1 _{OCR}	Zero-shot Learning	4.37 → 4.59	21.42 → 21.89	1.50 → 1.72	34.06 → 40.24
	QLoRA Fine-tuning	12.11 → 12.33	30.52 → 30.78	2.77 → 2.98	53.19 → 60.38
	Full Fine-tuning	12.42 → 12.75	31.72 → 32.19	3.32 → 3.38	65.84 → 67.53
Qwen2-Audio	Zero-shot Learning	4.29 → 4.51	21.53 → 22.18	1.59 → 1.77	34.31 → 40.52
	QLoRA Fine-tuning	12.05 → 12.19	30.77 → 31.04	2.80 → 3.01	54.27 → 61.44
	Full Fine-tuning	12.37 → 12.68	31.63 → 32.12	3.22 → 3.25	66.62 → 68.25

Table 5.10: Performance of baseline vs. planning models in non-video settings across different learning regimes. Each cell shows the result *before* → *after* applying the planning method.

Notably, despite these gains, video-based planning models such as Plan-mPLUG-Owl3 still outperform their non-visual counterparts by a notable margin. Nonetheless, our findings reinforce the idea that structured planning improves summarization quality beyond the video domain.

Human Evaluation. We conduct a human evaluation on 50 randomly selected instances from the VISTA test set. Annotators include master’s and doctoral students

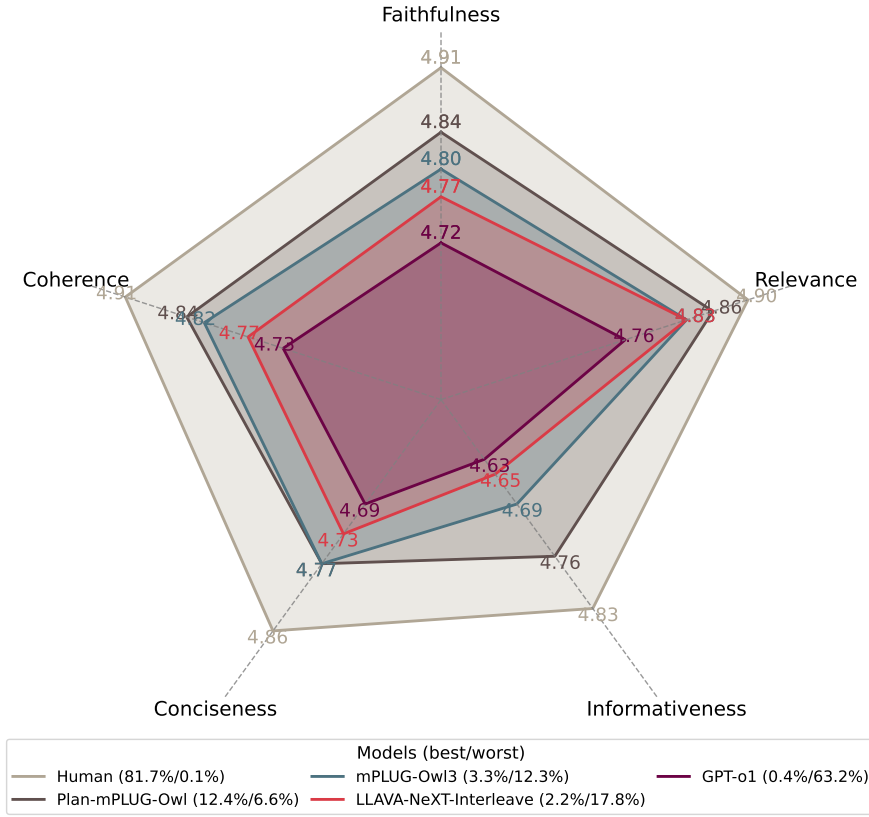


Figure 5.7: Human evaluation results. Human-written summaries consistently outperform all neural models.

in computer science or computational linguistics with advanced English proficiency. They receive compensation per our university’s standard rate and are blind to the source of each summary to ensure impartial assessment. We compare Plan-mPlug-Owl3, mPLUG-Owl3, LLaVA-NeXT-Interleave, and GPT-o1 against human reference summaries. Three independent annotators are asked to review the source video and evaluate corresponding model outputs (and the human upper bound) on a 1–5 Likert scale for Faithfulness, Relevance, Informativeness, Conciseness, and Coherence (higher scores indicate better quality). They are also asked to provide an overall ranking. In total, participants rated 750 samples ($50 \times 5 \times 3$). Appendix Section C.3 contains the full evaluation instructions.

Figure 5.7 presents the performance of each model, along with the proportion of instances where models are rated best or worst. Fleiss’ Kappa scores for Faithfulness ($\kappa = 0.767$), Relevance ($\kappa = 0.842$), Informativeness ($\kappa = 0.721$), Conciseness ($\kappa =$

0.792), and Coherence ($\kappa = 0.813$) indicate a substantial level of agreement, with an average agreement score of $\kappa = 0.787$. Overall, human-written summaries outperform all neural summarization models in quality, as they are perceived as substantially more faithful, coherent, concise, and informative. Human-written summaries are 81.7% more likely to be rated as best compared to model-generated summaries.

Among the four neural models, GPT-o1 performs worst, being rated as worst 63.2% of the time. LLAVA-NeXT-Interleave follows suit, with a 17.8% chance of receiving the worst ranking. The plan-based model, Plan-mPLUG-0w13, outperforms mPLUG-0w13 and demonstrates superior performance across all metrics. Additionally, it stands out among neural summarization systems for its higher likelihood of generating high-quality summaries. Paired t-tests show that human answers are considered significantly better than all neural models in all metrics ($p < 0.05$), revealing a clear gap between automatic systems and human performance on the VISTA dataset. The plan-based method is significantly better ($p < 0.05$) than other neural models in faithfulness, coherence, and informativeness, although it falls short of human performance.

LMM-as-a-Judge Evaluation. To facilitate large-scale comparisons of model outputs, we adopt a method inspired by LLM-as-a-judge (Liusie et al. 2024; Liu et al. 2024d; Zheng et al. 2023; Liu et al. 2025b), extending it to use a large multimodal model (Chen et al. 2024b). The proposed LMM-based evaluator incorporates both textual and video modalities and assesses the same summary quality dimensions used in our human evaluation for trend-level comparisons. Specifically, we use GPT-o1 as the evaluator. To minimize potential bias from prior queries, the conversation history is reset before each evaluation. The instructions for the LMM-as-a-Judge evaluation are provided in Appendix Section C.4.

We validate the agreement between GPT-o1 and human ratings by comparing its ratings with human evaluations on the same 50 samples from the VISTA test set. We calculate Fleiss' Kappa between GPT-o1 and mean human ratings across the dimensions of Faithfulness ($\kappa=0.732$), Relevance ($\kappa=0.803$), Informativeness ($\kappa=0.730$), Conciseness ($\kappa=0.792$) and Coherence ($\kappa=0.721$) at instance level. These results indicate that human evaluators and GPT-o1 achieve substantial levels of agreement across these dimensions. Following this, we expand the evaluation to include all samples in our test set.

Compared to fine-tuned models, GPT-o1 assigns the lowest scores to its own re-

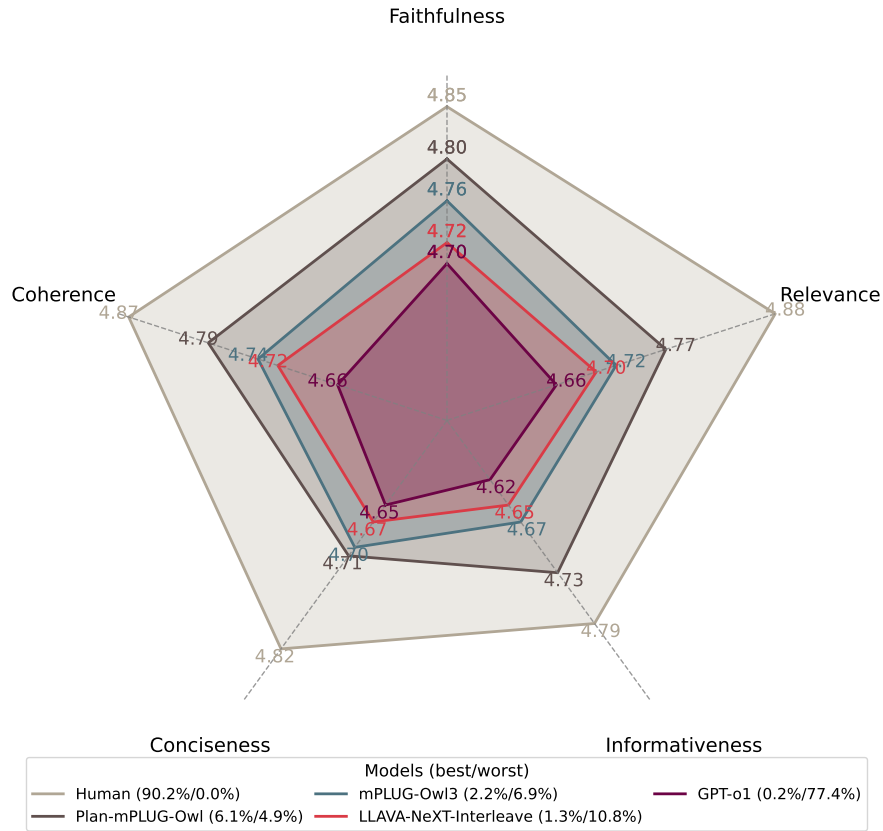


Figure 5.8: LMM-as-a-Judge evaluation results showing that human-written summaries consistently outperform neural models.

sponses (see Figure 5.8). Human-written summaries consistently receive the highest scores and are generally regarded as the best. Aligning with our human evaluations, GPT-o1 also recognizes that the plan-based model outperformed other models. We further conduct paired t-tests to find that human summaries outperform all neural models across all metrics with statistical significance ($p < 0.05$). Moreover, the plan-based model demonstrates significantly better performance ($p < 0.05$) than other neural models across all metrics except for conciseness. Our results also indicate that although the plan-based method can improve the performance of end-to-end models to some extent, there is still a gap between machine-generated and human summaries, which also reflects the challenging nature of our dataset.

Case Study and Error Analysis. For our case study, we randomly select a sample (Kübler et al. 2020) from the test split. The analysis in Table 5.11 reveals differences

in summary quality across models and against the human-written text. Specifically, GPT-o1 often produces concise summaries but at the cost of precision. For example, it incorrectly claims that “data splitting helps control test thresholds,” which is a hallucination — while data splitting ensures a tractable null distribution, it does not explicitly control test thresholds. Furthermore, its summaries frequently oversimplify complex concepts, reducing the depth of explanations and omitting crucial distinctions, such as the role of dependency calibration in the proposed method. Similarly, mPLUG-0w13 introduces factual inaccuracies, such as stating that data splitting “ensures a reliable null distribution.” This phrasing misleadingly implies that reliability is an inherent property of data splitting, whereas the correct point is that it makes the null distribution tractable rather than necessarily more reliable.

Plan-mPlug-0w13 is more factually accurate than the other models. It correctly captures the main idea of full-sample hyperparameter learning and testing without data splitting. However, it still introduces subtle distortions, such as falsely suggesting a “trade-off” between test power and tractability, which misrepresents the actual relationship. These inaccuracies, while less severe than those in GPT-o1 and mPLUG-0w13, highlight the model’s tendency to infer unstated causal links, leading to potential misinterpretations. Despite the relative strengths of Plan-mPlug-0w13, all generated summaries fall short of human-written text. The model-generated outputs consistently struggle with informativeness, coherence, and factual accuracy. These shortcomings underscore the ongoing challenge of improving automated summarization systems to better align with human standards in both accuracy and clarity.

Controlled generation experiments reveal that hallucination issues are further amplified when imposing constraints on readability and length. Under readability control (Table 5.12), GPT-o1 is more likely to introduce fabricated or misleading content when forced to generate more complex outputs. This occurs because it lacks an explicit mechanism to ensure factual consistency while adapting to varying readability demands. Rather than relying on implicit internal heuristics, Plan-mPlug-0w13 has an explicit planning mechanism, which makes it less likely to introduce unsupported claims. Planning provides an additional layer of control, helping the model maintain factual alignment even as readability demands change. A similar trend is observed in length control experiments (Table 5.13). As the compression ratio increases, GPT-o1 struggles to balance conciseness and informativeness, sometimes hallucinating missing details to compensate for omitted content. This suggests that purely instruction-based control

(i.e., prompting the model to shorten outputs) does not effectively enforce content retention, leading to greater inconsistencies. In contrast, the plan allows Plan-mPlug-Owl3 to selectively retain essential elements, reducing the risk of generating misleading content; it can also avoid answering deleted questions, to a certain extent.

These findings reinforce the advantages of plan-based control over instruction-based prompting. While neither approach fully eliminates hallucinations, planning provides a structured mechanism to manage content selection, ensuring greater alignment with the input source compared to freeform generative adjustments.

Model	Output
Reference	Modern large-scale kernel-based tests such as maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD) optimize kernel hyperparameters on a held-out sample via data splitting to obtain the most powerful test statistics. While data splitting results in a tractable null distribution, it suffers from a reduction in test power due to a smaller test sample size. Inspired by the selective inference framework, we propose an approach that enables learning the hyperparameters and testing on the full sample without data splitting. Our approach can correctly calibrate the test in the presence of such dependency and yield a test threshold in closed form. At the same significance level, our approach's test power is empirically larger than that of the data-splitting approach, regardless of its split proportion.
GPT-o1	Modern kernel-based tests like MMD and KSD use data splitting to tune hyperparameters. This splitting helps control test thresholds but reduces power because it splits the data into smaller parts. We propose a new method to tune hyperparameters without splitting the data. Our method provides better test power than data splitting.
mPLUG-0w13	Large-scale kernel-based tests like maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD) usually split the data to tune hyperparameters. While this approach ensures a reliable null distribution , it reduces test power due to smaller sample sizes. We propose a method that avoids data splitting, enabling full-sample hyperparameter learning and testing. This method improves test power at the same significance level compared to data splitting, with a closed-form threshold for the test.
Plan Questions	q1: How do kernel-based tests like MMD and KSD typically optimize hyperparameters? q2: What is the key idea behind the proposed approach? q3: What are the advantages of the proposed method compared to data-splitting methods?
Plan-mPlug-0w13	Modern kernel-based tests like maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD) typically split the data to optimize hyperparameters, trading off test power for a tractable null distribution. Inspired by the selective inference framework, we propose a novel approach that eliminates the need for data splitting, allowing full-sample hyperparameter learning and testing. Our method addresses dependencies introduced by learning and provides a closed-form test threshold while achieving higher test power than data-splitting methods, regardless of split proportions.

Table 5.11: Comparative case study of model outputs. Hallucinations are highlighted in red.

Model	Output
GPT-o1	Modern kernel-based tests, such as MMD and KSD, use data splitting to adjust hyperparameters . This data splitting ensures proper test thresholds but weakens test power because the data is divided into smaller portions. We propose a new method that avoids splitting the data and provides better test power.
Plan Questions	q1: What do MMD and KSD do with data splitting? q2: What is the main purpose of the proposed method? q3: Why is the proposed method better than data splitting?
Plan-mPlug-0w13	Kernel-based tests like maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD) typically divide the data into subsets to adjust hyperparameters . This approach ensures a manageable null distribution but reduces test power. Inspired by selective inference, we propose a new method that avoids splitting the data. This method uses the entire dataset for hyperparameter tuning and testing, textcolorredeliminating dependencies introduced by learning. It also provides a clear, closed-form test threshold and achieves higher test power compared to data-splitting methods, regardless of the split proportions.

Table 5.12: Decreasing reading difficulty: plan modifications vs instruction adjustments. Hallucinations are in red.

Model	Output
GPT-o1	Kernel-based statistical methodologies, exemplified by maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD), conventionally employ data partitioning mechanisms to facilitate hyperparameter optimization. While this stratagem ensures a well-calibrated null distribution , it inherently compromises the statistical power of the test due to data fragmentation. We advance a paradigm wherein hyperparameter selection is conducted holistically , circumventing data splitting and consequently preserving statistical power.
Plan Questions	q1: Through what operational principles do MMD and KSD navigate the interplay between hyperparameter tuning and null distribution calibration? q2: How does the selective inference framework inform the methodological innovations of the proposed approach? q3: What are the quantifiable and theoretical advantages of the proposed strategy in circumventing the limitations imposed by data-splitting conventions?
Plan-mPlug-0w13	Modern kernel-based tests like maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD) typically split the data to optimize hyperparameters, trading off test power for a tractable null distribution. Inspired by the selective inference framework, we propose a novel approach that eliminates the need for data splitting, allowing full-sample hyperparameter learning and testing. Our method addresses dependencies introduced by learning and provides a closed-form test threshold while achieving higher test power than data-splitting methods, regardless of split proportions.

Table 5.13: Increasing reading difficulty: plan modifications vs. instruction adjustments. Hallucinations are in red.

Model	Output
GPT-o1	Kernel-based tests like MMD and KSD use data splitting for hyperparameter tuning, which controls thresholds but reduces power. We propose a method that avoids splitting, improving test power .
Plan Questions	q1: How do kernel-based tests like MMD and KSD typically optimize hyperparameters? q2: What are the advantages of the proposed method compared to data-splitting methods?
Plan-mPlug-0w13	Kernel-based tests, such as maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD), rely on data splitting for hyperparameter tuning. Our method delivers analytically derived test thresholds and consistently outperforms data-splitting techniques in test power, regardless of how the data is divided.

Table 5.14: Deleting 30% of plan questions vs. instruction adjustments. Hallucinations are highlighted in red.

5.6 Related Work

Video-to-Text Summarization. Video-to-Text Summarization generates coherent summaries by integrating multimodal information (Hua et al. 2025), supported by datasets like MSS (Li et al. 2017), VideoXum (Lin et al. 2024b), MMSum (Qiu et al. 2024), Hierarchical3D (Papalampidi et al. 2023), and LfVS-T (Argaw et al. 2024), spanning tasks from instructional videos to general web content (Li et al. 2017; Zhou et al. 2018; Li et al. 2020a; Li et al. 2020b; Liu et al. 2021a; Fu et al. 2021; Liu et al. 2022a; Krubiński et al. 2023; Han et al. 2025b; He et al. 2023; Hua et al. 2025; Islam et al. 2024a; Qiu et al. 2024). Technical advancements include hierarchical attention models (Sanabria et al. 2018), extractive methods using multimodal features (Cho et al. 2021; Krubiński et al. 2023), and hybrid extractive-abstractive frameworks (Ramakrishnan et al. 2022; Papalampidi et al. 2023). Transformer-based systems have further improved performance (Krubinski et al. 2023; Li et al. 2020b; Shang et al. 2021; Mahon et al. 2024a). However, challenges in summarizing academic videos remain under-explored.

Scientific Text Summarization. Scientific Text Summarization condenses complex scholarly content into concise formats (Cachola et al. 2020; Ju et al. 2021; Liu et al. 2023c; Liu et al. 2023b), supported by datasets like TalkSumm (Lev et al. 2019) for academic video transcripts, SumSurvey (Liu et al. 2024c) for survey papers, ACLSum (Takeshita et al. 2024) for ACL discourse, and SciNews (Liu et al. 2024b) for simplifying research for broader audiences. M³AV (Chen et al. 2024c) supports tasks like ASR, TTS, and slide-script generation. Methods like RST-LoRA (Liu et al. 2024a) and RSTformer (Liu et al. 2023c) improve discourse and structural summarization, while CiteSum (Mao et al. 2022b) and SSR (Fatima et al. 2023) focus on scalability and audience-specific customization. Despite these efforts, scientific summarization remains a challenging domain due to the inherent complexity and diversity of scholarly texts.

Plan-based Summarization. Plan-based Summarization employs structured representations to improve summary quality and reduce hallucinations (Narayan et al. 2021; Amplayo et al. 2021; Wang et al. 2022; Narayan et al. 2023; Liu et al. 2025b). Research focuses on text-based planning with elements like entities (Narayan et al. 2021; Liu et al. 2021d; Huot et al. 2024), keyword prompts (Creo et al. 2023), and question-answer pairs (Narayan et al. 2023). Examples include PlanVerb (Canal et al. 2022), which con-

verts task plans into natural language via semantic tagging, and domain-specific approaches that align with knowledge structures for improved quality (Srivastava et al. 2024). Blueprint-based frameworks utilize intermediate plans to create coherent narratives for visual storytelling (Liu et al. 2023a). However, plan-based strategies for multimodal tasks, particularly video-to-text summarization, have received limited attention.

5.7 Chapter Summary

This chapter addressed Research Question 3 by investigating whether discourse-aware planning can support macro-level organization and factual grounding when the input evidence is multimodal and temporally fragmented. To enable systematic evaluation, the chapter first introduced VISTA, a dataset of 18,599 aligned scientific presentation videos and paper abstracts. Benchmarking across proprietary and open-source multimodal large models confirmed that scientific video-to-text summarization remains a challenging task, with end-to-end generation often struggling to reproduce the organizational conventions of scientific abstracts. Building on the discourse-conditioning theme developed in the preceding chapters, we tested whether a planning strategy that treats a summary as a series of answers to ordered questions, grounded in the QUD framework, could help reduce these organizational shortcomings. The results indicated that explicit planning tends to improve summary quality, factual consistency, and alignment with video content across zero-shot, QLoRA, and full fine-tuning settings, providing support for Hypothesis III.

These findings suggest that separating organizational decisions from surface generation through an explicit planning stage can be beneficial even when the input is no longer a coherent text document. However, the planning mechanism explored in this chapter is primarily concerned with the macro-level structure of expert-style abstracts, where the rhetorical moves are relatively predictable. The next chapter shifts to lay summarization, where the target audience changes and the summary must additionally incorporate explanatory content whose placement, scope, and discourse function require more fine-grained control.

6

Discourse-Aware Planning for Explanatory Summarization

Chapter 3 through Chapter 5 progressively develop a discourse-guided summarization framework under increasingly complex conditions. Chapter 3 transforms RST relations into type-aware probabilistic distributions and injects them into sparse attention encoders, which supports Research Question 1 (RQ1) and Hypothesis I. Chapter 4 extends this approach under parameter-efficient fine-tuning constraints and provides empirical support for Research Question 2 (RQ2) and Hypothesis II. Chapter 5 advances structured generation to multimodal settings by introducing intermediate planning signals based on the Question under Discussion (QUD) framework, consistent with Research Question 3 (RQ3) and Hypothesis III. A shared insight across these chapters is that modeling organizational decisions as explicit intermediate variables leads to more controllable and diagnostically useful improvements. Building on this foundation, Chapter 6 turns to a task more closely aligned with public communication, namely the generation of lay summaries of scientific content for non-expert audiences. Unlike expert-oriented abstracts, lay summaries should not only compress information but also provide explanatory content that reduces barriers to understanding through background elaboration, analogies, or causal reasoning. The challenge is that existing end-to-end models typically do not explicitly model where explanations are needed, what their referents are, or what discourse function they should serve. As a result, the generated summaries contain too few explanatory sentences or fail to connect them clearly with the main content. Against this backdrop, Research Question 4 (RQ4) asks how explanatory content in public-facing

summarization can be treated as a controllable discourse element to improve accessibility and readability without compromising factual reliability.

To this end, **Chapter 6** tests Hypothesis IV that, when modeled as controllable discourse elements, explanations can enhance the accessibility and readability of public-facing summaries while preserving factual reliability, provided that their discourse scope, target specificity, and placement are explicitly managed. Concretely, **Chapter 6** integrates explanatory modeling into the planning-based framework developed in earlier chapters by using an RST parser to extract explanatory elementary discourse units (EDUs) and their corresponding target EDUs from reference lay summaries. These explanations are then treated as answers to implicit questions based on the QUD framework, which are made explicit as ordered plans. Two model variants are trained, one that takes the plan as input conditions (Plan-Input) and another that uses it as a prefix to guide generation (Plan-Output). To better align evaluation with task requirements, **Chapter 6** also introduces adjustments on the assessment side. This evaluation strategy incorporates external knowledge verification through retrieval and fact-checking methods based on sources such as Wikipedia, alongside human evaluations to assess factual correctness and the usefulness of explanations. Subsequent experiments on three lay summarization datasets (SciNews, eLife, and PLOS) evaluate whether the discourse-driven planning approach can yield measurable gains in explanation quality and factual consistency. **Chapter 6** also compares plan-based variants against end-to-end baselines and analyzes their relative impact on accessibility, readability, and the proportion of explanatory content in generated summaries.

6.1 Problem Formulation

In the domain of lay summarization for scientific documents, the inclusion of explanatory content is beneficial for improving readability and accessibility, particularly when dealing with difficult concepts or complicated statements (Srikanth et al. 2021; August et al. 2022; Luo et al. 2022; Goldsack et al. 2022). As an example, consider the target sentence being explained (highlighted in orange) within the lay summary in **Figure 6.1**. It describes the role of the cerebellum (a similar sentence can be found in the source document, also highlighted in orange) and is accompanied by an explanation highlighted in green, which compares the cerebellum to a coach. This analogy, which illustrates how the cerebellum adjusts the timing of actions based on previous movements, makes

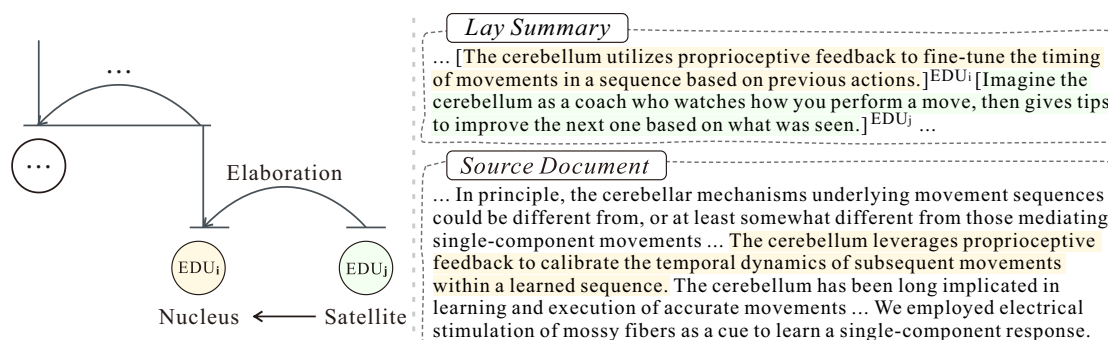


Figure 6.1: An excerpt of a source document paired with its summary (bottom); the explanatory sentence and its target are highlighted in green and orange, respectively. The RST tree (top) for the text corresponding to the two sentences shows they are linked by the rhetorical relation *Elaboration*.

the preceding technical description easier to understand. Using simpler vocabulary or shorter sentences can make the summaries easier to read, but could also lead to misinterpretations (Liu et al. 2023b; Han et al. 2023; Wu et al. 2023b; Hewett et al. 2024). Thus, explanations are often used to avoid this, balancing accessibility with accuracy and cognitive load (Oksa et al. 2010).

Empirical analysis of lay summaries supports the observation that explanations are commonplace. Specifically, by using a discourse parser (Liu et al. 2020; Liu et al. 2021e) to identify explanations in the eLife and PLOS (lay summarization) datasets (Goldsack et al. 2022), we find that explanations account for approximately 19.02% and 18.19% of summary sentences, respectively. This proportion of explanations in lay summaries is three times higher than in expert summaries (6.28% and 5.06%) and nearly four times greater than in the original academic papers (5.16% and 4.76%). In other words, explanations are significantly more common than other discourse relations like *Condition* and *Purpose*. This underscores their importance in structuring lay summaries and enhancing text accessibility.

Most lay summarization models follow an end-to-end approach (Goldsack et al. 2022; Liu et al. 2022a; Goldsack et al. 2023; Liu et al. 2024b) without *explicitly* accounting for explanations. As a result, explanations are often underrepresented in generated summaries, which in turn may explain why generated summaries lack the clarity and readability found in human-produced ones (Guo et al. 2021b; Goldsack et al. 2023; Tang et al. 2023a; Liu et al. 2023b; Zhang et al. 2024c; Wang et al. 2025a). In this chapter, we develop neural models that are capable of generating *lay summaries* with controlled ex-

planatory content. We achieve this by *planning* the content of the summary, thus *directly* steering the model towards generating explanations. Plan-based models are a suitable type of approach for lay summarization, as they are less prone to hallucinations, offer greater controllability (Puduppully et al. 2019; Narayan et al. 2021; Moryossef et al. 2023; Narayan et al. 2023; Huot et al. 2023; Huot et al. 2024; Liu et al. 2025a), and can be flexibly adapted to different generation tasks depending on how plans are formulated.

Our method conceptualizes plans as a series of questions that trigger explanatory content, with their strategic placement learned from training data. This placement is not arbitrary; rather, it reflects where human writers tend to insert explanations to clarify complex claims or findings. Importantly, our approach not only models where to explain, but also what is being explained and how—e.g., through analogy, background, or causal reasoning. In addition, evaluating such explanations remains challenging, as automatic consistency detection metrics often misclassify useful additions as hallucinations (Cao et al. 2022a). To address this, we incorporate external knowledge verification, but acknowledge that finer-grained human or LLM-based judgments are still needed for nuanced evaluation.

A common challenge with plan-based models is the scarcity of datasets with plan annotations. As a result, plans are often reverse-engineered from existing reference summaries through automatic means, e.g., by identifying entities (Narayan et al. 2021; Huot et al. 2024) or generating questions and their answers (Narayan et al. 2023). Our approach draws inspiration from Rhetorical Structure Theory (RST; Mann et al. 1987). RST represents texts as trees (see Figure 6.1), where the leaves correspond to Elementary Discourse Units (EDUs) and the nodes specify how these and larger units are connected through rhetorical relations (e.g., *Elaboration*). Discourse units are further characterized in terms of their text importance: *nuclei* represent central segments, whereas *satellites* denote peripheral ones. As depicted in the right part of Figure 6.1, EDU_j acts as a satellite to EDU_i , supplying an explanation.

We automatically generate plans from reference summaries annotated with RST trees (Liu et al. 2020; Liu et al. 2021e). We hypothesize that explanatory EDUs answer latent questions (Beaver et al. 2017), which we verbalize. Our plans are designed to pose these latent questions, but we do not state the answers explicitly (see Figure 6.2). We develop two model variants that differ in how they integrate planning in the summarization process, with plans either included as part of the source document or as part of the target summary. Empirical results using Mistral (Jiang et al. 2023a) as backbone confirm

that our discourse-driven planning approach significantly improves the quality of lay summaries but also has the potential to control the generation of explanatory content to a certain extent.

6.2 Methodology

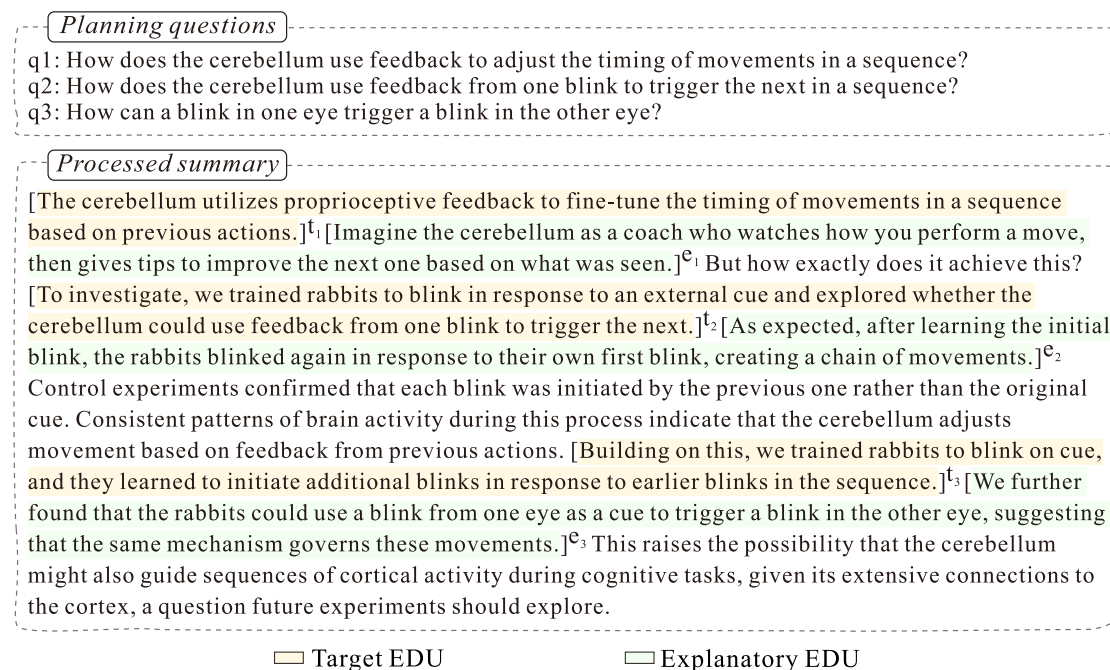


Figure 6.2: We use DMRST (Liu et al. 2021e) to extract explanatory (*e*) EDUs and their target (*t*) EDUs from reference summaries, and then feed this data into GPT-4o to generate plans (*b*).

Task Definition. We formalize *explanatory summarization* as follows:

$$\text{Let } D = \{(x_1, s_1), (x_2, s_2), \dots, (x_n, s_n)\} \quad (6.1)$$

denote a dataset where each tuple (x_i, s_i) contains a document x_i and its corresponding lay summary s_i . We further assume s_i contains some explanatory elements e_i such as examples, contextual explanations, or background information. Our objective is to devise a summarization model that can model the generation of this explanatory content.

Model Overview. Inspired by the blueprint approach (Narayan et al. 2023), we propose a discourse-driven planning method to achieve the above goal. This involves transforming the original dataset D into

$$D' = \{(x_1, b_1, s_1), (x_2, b_2, s_2), \dots, (x_n, b_n, s_n)\} \quad (6.2)$$

by introducing a planning component b . We then train a summarization model to learn the conditional probability distribution $P(S|X, B)$ or $P(B, S|X)$, where $S = \{s_1, \dots, s_n\}$, $X = \{x_1, \dots, x_n\}$, and $B = \{b_1, \dots, b_n\}$. Each b_i in B consists of a series of ordered plan questions (q_1, q_2, \dots, q_n) , which do not initially exist in the original dataset D .

Note that $P(S|X, B)$ assumes plan b is treated as part of the input sequence, whereas $P(B, S|X)$ treats b as a prefix to summary generation. The key challenge in training either model is that plan b is latent. We thus automatically augment the original training pairs (x, s) into triplets (x, b, s) , with b acting as a conditional or control signal, offering an overarching plan for summary generation while also guiding the creation of explanatory sentences by prompting responses to the plan.

Explanatory Content Extraction. To obtain triplets (x, b, s) , we apply DMRST (Liu et al. 2020; Liu et al. 2021e), an RST-based parser, to automatically identify and extract all EDUs functioning as explanations in reference summary s (see Figure 6.2). Out of 18 rhetorical categories, we select four that are relevant to explanatory sentences: *Background*, *Elaboration*, *Explanation*, and *Comparison* (Table 6.1 has the full list). Given the directed nature of RST relations, we can identify discourse units that serve as explanations and the content they explain (see Figure 6.1). Note that the extracted EDU pairs act only as a silver standard proxy for the subsequent training phase of our models.

RST type	RST label
<i>Background</i>	Background, Circumstance
<i>Elaboration</i>	Elaboration-additional, Elaboration-general-specific, Elaboration-part-whole, Elaboration-process-step, Elaboration-object-attribute, Elaboration-set-member, Example, Definition
<i>Explanation</i>	Evidence, Explanation-argumentative, Reason
<i>Comparison</i>	Comparison, Preference, Analogy, Proportion, Topic-Comment

Table 6.1: RST explanatory relations.

Explanatory Plan Generation. We propose that explanatory sentences can serve as responses to (implicit) questions raised by preceding target sentences, aligning with the theoretical framework of Question under Discussion (QUD). We employ GPT-4o (OpenAI, 2023) to generate these questions from target sentences (attested in reference summary) and the preceding context, as illustrated in Figure 6.2. For example, the plan question q_2 is generated based on the target sentence t_2 and the sentences that precede it. We do not filter the generated questions, but we create the plan by following the order in which explanations appear in the reference. The prompt for generating plan questions can be found in Appendix Figure D.1.

Explanatory Summarization Models. We propose two model variants. The first variant (Plan-Output) concatenates all plan questions b with the summary s to form a sequence $[b; s]$ for each document-summary pair. The training objective for this model is to generate both the plan and the summary based on the input document x .

The second variant (Plan-Input) is not an end-to-end model. Instead, it employs a Plan Generation (PG) module and a Summary Generation (SG) module, which are trained separately. The PG module learns to generate plan b from the input document x , and is trained on (x, b) pairs, where x is the input and b is the silver standard plan. The SG module is trained on $([x; b], s)$ tuples where $[x; b]$ represents the input document enriched with plan b (we concatenate input document x with plan b). At test time, the trained PG module is used to predict plan \hat{b} based on input document x . Generated questions \hat{b} are then combined with x to form the input $[x; \hat{b}]$ for the SG module, which produces the final summary on the test set.

Plan-Output trains a single end-to-end model that learns to *jointly* generate the plan and the summary. This design reduces inference overhead, making it well-suited for real-time applications or resource-constrained environments (since only one inference call is made during testing). In contrast, Plan-Input adopts a modular approach, training the PG and SG components *separately*. This makes Plan-Input better suited to settings where task requirements evolve over time, allowing for the refinement or replacement of individual module components as needed.

Note that the original Blueprint approach of Narayan et al. (2023), relies on *phrase-level* control (the answers to the questions are mostly named entities) and, as such, cannot be readily used to control the generation of explanations that are typically longer and semantically richer. Moreover, the frequent repetition of entities (i.e., the same entity is

often mentioned multiple times) adds unnecessary complexity to the planning process, necessitating the use of filtering algorithms to streamline it. In contrast, planning at the *EDU-level* provides a more cohesive structure and reduces redundancy.

6.3 Experimental Setup

Datasets. We report results on three lay summarization datasets, namely SciNews (Liu et al. 2024b), eLife (Goldsack et al. 2022), and PLOS (Goldsack et al. 2022).

Automatic Evaluation Metrics. We report several complementary metrics aimed at assessing different facets of summary quality. We use Rouge (Lin. 2004) and BERTScore (Zhang et al. 2020b) to evaluate *informativeness* against the human references; we report F1 of Rouge-2 (R2) and Rouge-Lsum (RLsum). We use D-SARI (Sun et al. 2021) and Flesch Reading Ease Formula (FRE, Kincaid et al. 1975) to measure summary *readability*. D-SARI and FRE scores range from 0 to 100, with higher values indicating easier-to-read material and lower scores reflecting more complex passages. Additionally, we compute `ExpRatio`, which represents the *proportion of explanatory* EDUs to the total number of EDUs in the generated summary (explanatory and target EDUs always appear in pairs due to the RST structure). A higher `ExpRatio` suggests the summary contains more explanations. We also report the generated summary *length*, which we calculate as the average number of summary tokens (AST) using `spaCy`.

We assess *factual consistency* between the source document and the generated summary using `SummaCConv` (Laban et al. 2022), which determines whether summary sentences are entailed by the input. However, human-written lay summaries often include additional external information, which could be mistakenly classified as *hallucinations* when assessed automatically with original `SummaCConv`. This highlights the limitations of traditional consistency detection metrics for our task, where only extrinsic hallucinations, rather than all additional explanations, should be penalized.

To address this issue, we propose a new `SummaC` variant. Specifically, for summary sentences with entailment scores below 0.5 (i.e., not supported by the source text), we use the `Wikipedia-API` to retrieve relevant articles and re-evaluate these sentences using `SummaCConv`. This approach allows us to determine whether sentences unsupported by the source text can be validated using external knowledge bases. If the highest entailment score from the retrieved articles is higher than the original score, it replaces the original

score for that sentence. We report both the original SummaC_{Conv} and the proposed variant, which we call SummaC^* . We also leverage VeriScore (Song et al. 2024b) to verify whether the claims in the generated summary align with objective facts by consulting external knowledge bases. For claim extraction and verification, we use GPT-4o.

Model Comparisons. We build Plan-Output and Plan-Input (PG and SG) on top of the $\text{Mistral-7B-Instruct-v0.3}$, which we fully fine-tune on the above datasets (enriched with plan annotations). We compare our models against the following Mistral configurations: a) zero-shot setting (Mistral_{ZS}); b) in-context learning with one randomly selected demonstration from the training split (Mistral_{ICL}); and c) full parameter fine-tuning without planning (Mistral_{FT}).

We also re-implement the best-performing multi-task blueprint model (Blueprint_{MT}) from Narayan et al. (2023), which is optimized for two tasks: 1) question generation given source document and said answer, or 2) summary generation given the same document and its question-answer plan. We use GPT-4o to identify answers (while the original Blueprint model extracts answers using spaCy , we employ GPT-4o which leads to better performance – on average 2.21% higher RLsum across datasets) and generate corresponding questions and fine-tune with the same backbone Mistral model on datasets enriched with question-answer plans.¹

We also include results with GPT-4o (OpenAI. 2023) in zero-shot (GPT-4o_{ZS}) and in-context learning (GPT-4o_{ICL}) settings; we compare against the state-of-the-art (SOTA) methods on each dataset (results are directly taken from respective publications). For a fair comparison, we use the same prompt and/or selected sample for both $\text{Mistral}_{ZS/ICL}$ and $\text{GPT-4o}_{ZS/ICL}$. After inference, we discard all generated plans, retaining only the summaries for automatic evaluation.

Hyper-parameter Settings. For all fine-tuning experiments, we utilize the AdamW optimizer (Loshchilov et al. 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$, and weight decay = 0.1, along with a warm-up ratio of 0.15. The initial learning rate is set to $5e-5$, employing cosine learning rate scheduling. We set the random seed to 2,024 and the dropout rate to 0.1. For SciNews, we set the maximum number of new generation tokens to 1,024, for eLife to 512, and for PLOS to 256. All other parameters follow the default settings of the Transformers library. During training, we save the checkpoint with

¹ Prompts are offered in Appendix Figure D.2, Figure D.3, Figure D.4 and Figure D.5.

the highest Rouge-2 F1 on the validation set as the final checkpoint. All experiments run for 15 epochs with a batch size of 8, and early stopping is implemented to prevent overfitting, with all models converging before reaching 15 epochs. During inference, we employ beam search with size 4, length penalty of 3.0, and set the no-repeat n-gram size to 3.

For GPT-4o,² we use version *gpt-4o-2024-05-13* (our experiments took place May 1, 2024, to March 31, 2025). We set temperature to 1, top_p to 1, frequency penalty to 0.2, and presence penalty to 0.2. All other hyperparameters follow OpenAI’s default values.

6.4 Results and Analysis

Our results are summarized in [Table 6.2](#), which consists of three main blocks corresponding to different datasets. Within each block, we compare models in zero-shot and in-context learning settings against fine-tuned systems and, where applicable, against the human reference summaries.

We find the performance of fine-tuned models to be superior. Across all metrics and datasets, Plan-Input emerges as the top-performing model, followed closely by Plan-Output (see R2 and RLsum). Recall that Plan-Input is a pipeline system where the plan and summary generation stages are trained separately, which we hypothesize makes the summarization task easier.

In comparison to the related `BlueprintMT` and other comparison models, our two discourse-driven systems produce summaries that are easier to read (see FRE and D-SARI), contain more explanations (see `ExpRatio`), are closer to human-level performance (see Reference Summary), and align more closely with human summaries in terms of length (AST). Since `BlueprintMT` relies on entities and phrases to generate plans, it lacks diversity in the types of questions it predicts; in fact, *what* and *who* questions are the greatest majority with an average of 82.2% across datasets. In contrast, our models produce a more balanced distribution, with *what* and *who* questions making up 23.2% of the plan, while *how* and *why* questions account for an average of 66.4%.

As human-authored summaries often include content that is not directly mentioned in the source document ([Cao et al. 2022a](#)), we observe relatively low (original) SummaC scores. However, when external knowledge bases are considered, we find that human

² <https://platform.openai.com/docs/models/>

Data	Model	R2	RLsum	BERTscore	D-SARI	FRE	ExpRatio	AST	SummaC/SummaC*	VeriScore
SciNews	Mistral _{ZS}	7.07	37.02	57.21	14.12	27.41	10.45	554.29	46.11/52.65	0.43
	Mistral _{ICL}	7.11	37.31	57.57	14.65	31.93	9.87	602.33	50.12/59.27	0.48
	GPT-4o _{ZS}	12.79	40.51	58.15	20.15	33.05	11.73	611.09	58.21/64.96	0.56
	GPT-4o _{ICL}	13.22	40.73	58.17	20.33	38.91	11.90	634.76	60.04/66.74	0.52
	Mistral _{FT}	15.56	46.12	64.38	30.11	38.25	13.61	669.27	64.10/69.73	0.56
	Blueprint _{MT}	15.62	46.10	64.33	32.35	40.01	15.03	688.93	72.35/75.18	0.62
	Plan-Output	15.73	46.30	65.34	36.23	42.25	17.51	685.33	72.29/75.03	0.67
	Plan-Input	15.88 ^{†‡}	46.41 ^{†‡}	65.32 ^{†‡}	37.18 ^{†‡}	43.11 ^{†‡}	17.68 ^{†‡}	692.14 ^{†‡}	72.40 ^{†‡} / 75.38 ^{†‡}	0.71 ^{†‡}
	Liu et al. (2024b)	14.92	45.19	62.80	—	—	—	—	—	—
	Reference Summary	—	—	—	—	42.23	18.91	694.80	45.08/81.94	0.81
eLife	Mistral _{ZS}	8.90	36.19	61.38	18.28	28.43	10.29	301.16	45.45/56.68	0.39
	Mistral _{ICL}	8.97	36.26	61.40	20.07	33.60	10.30	333.27	47.42/58.98	0.45
	GPT-4o _{ZS}	11.15	44.29	64.34	24.57	38.17	11.10	329.39	47.30/58.46	0.56
	GPT-4o _{ICL}	11.31	44.38	63.77	25.52	44.19	11.22	368.42	50.92/60.07	0.60
	Mistral _{FT}	14.40	47.68	86.87	32.60	48.42	13.56	377.59	58.31/62.33	0.58
	Blueprint _{MT}	14.85	48.09	87.22	35.31	50.33	14.51	375.47	60.31/65.44	0.62
	Plan-Output	14.91	48.27	87.57	38.77	56.76	17.71	377.61	61.33/68.21	0.70
	Plan-Input	15.11 ^{†‡}	48.64 ^{†‡}	87.92 ^{†‡}	39.22 ^{†‡}	55.48 ^{†‡}	17.73 ^{†‡}	380.39 ^{†‡}	61.35 ^{†‡} / 68.28 ^{†‡}	0.75 ^{†‡}
	Goldsack et al. (2023)	14.24	45.71	85.40	—	—	—	—	—	—
	Liu et al. (2024a)	14.92	48.21	87.81	—	—	—	—	—	—
Zhang et al. (2024c)	12.57	44.14	85.20	—	—	—	—	—	—	
Reference Summary	—	—	—	—	51.83	19.02	383.02	49.38/82.28	0.83	
PLOS	Mistral _{ZS}	6.43	31.42	63.24	19.09	29.70	11.03	130.14	45.83/67.47	0.41
	Mistral _{ICL}	6.50	31.66	63.31	21.27	35.93	11.19	149.30	49.02/68.87	0.44
	GPT-4o _{ZS}	11.57	35.39	64.22	21.06	32.69	12.31	155.25	55.23/70.79	0.52
	GPT-4o _{ICL}	11.72	35.40	64.50	23.74	39.44	12.31	152.38	58.02/73.03	0.50
	Mistral _{FT}	14.31	40.22	87.69	32.39	36.45	13.70	158.93	60.02/75.25	0.57
	Blueprint _{MT}	14.20	41.35	88.04	33.48	36.04	14.67	166.54	62.22/77.67	0.63
	Plan-Output	15.28	41.33	88.17	39.22	40.01	18.12	167.77	62.25/77.79	0.74
	Plan-Input	15.72 ^{†‡}	41.64 ^{†‡}	88.22 ^{†‡}	40.07 ^{†‡}	40.27 ^{†‡}	17.65 ^{†‡}	171.12 ^{†‡}	62.27 ^{†‡} / 77.95 ^{†‡}	0.72 ^{†‡}
	Goldsack et al. (2022)	13.52	38.63	—	—	—	—	—	—	—
	Zhang et al. (2024c)	12.33	40.60	85.70	—	—	—	—	—	—
Reference Summary	—	—	—	—	30.08	18.19	178.66	53.43/84.71	0.87	

Table 6.2: Model performance on three lay summarization datasets. Bold numbers represent the best results achieved by the models in each test set, excluding human results from the comparison. ExpRatio is the number of explanatory EDUs over all EDUs in summaries. Symbols [†] and [‡] denote that Plan-Input is statistically significant ($p < 0.05$) against Mistral_{FT} and Blueprint_{MT} using paired t-test, respectively.

summaries exhibit the lowest level of extrinsic hallucinations. VeriScore results further corroborate the observation that model-generated summaries lag behind human-authored ones regarding factual consistency. Nevertheless, fine-tuned models are overall more consistent, with plan-based models having an advantage against Mistral_{FT} and state-of-the-art systems. Both Plan-Input and Plan-Output achieve the highest consistency scores across datasets and metrics, indicating that planning at the EDU-level not only enhances consistency compared to phrase-level planning but also enables the generation of summaries with more elaborations while maintaining lower levels of hallucinations.

Impact of Plan Generation Ablation. Next, we ablate our plan generation strategy more closely, comparing it to naive alternatives. We apply Plan-Input as our main model, except when indicated otherwise.

1. Lead-3_Q and Lead-K_Q: Generate plans based on the first 3 or first K EDUs in the reference summary.
2. Tail-3_Q and Tail-K_Q: Generate plans based on the last 3 or last K EDUs in the reference summary.
3. Random-3_Q and Random-K_Q: Generate plans based on 3 random or K random EDUs in the reference summary.
4. All-EDUs_Q: Generate plans based on each of the EDUs in the reference summary.
5. NonExp-EDUs_Q: Generate plans based on each of the non-explanatory EDUs in the reference summary.
6. PG for zero-shot learning: Train a PG model and use the plan questions it generates as instructions to Mistral in a zero-shot setting.³

Here, K denotes the average number of explanatory EDUs parsed in training set summaries, rounded up to the nearest whole number ($K = 8$ for SciNews, $K = 4$ for eLife, and $K = 2$ for PLOS). For the first four experiments, we do not differentiate between explanatory and non-explanatory EDUs. Each selected EDU serves as the target sentence; all preceding EDUs are considered context and are used to generate the corresponding plan questions. For instance, in the All-EDUs_Q experiment (all sentences from the reference summary are involved in the plan generation, rather than being limited to explanatory EDUs), the first sentence is treated as a target EDU (since there are no preceding EDUs, there is no context in this case), and the first plan question is generated. Next, the second EDU is treated as a target EDU, and the first EDU is used as context, and so on.

As shown in Table 6.3, Lead-3_Q, Lead-K_Q, and All-EDUs_Q strategies demonstrate marginal improvements over Mistral_{FT} across all datasets, while all other strategies result in decreased performance. All strategies in Table 6.3 are worse than Blueprint_{MT}, Plan-Input, and Plan-Output, in terms of summary quality and factual consistency. This implies that heuristic ablations are less effective for the generation of plan questions. Results for NonExp-EDUs_Q and All-EDUs_Q also verify that modeling and generating plan questions for explanatory sentences is more valuable for lay summarization than

³ The prompt can be found in Appendix Figure D.5.

modeling and generating plans for other sentences. Results for `MistralZS+PG` further show that a robust plan generator improves performance, even in zero-shot settings.

Data	Model	R2	RLsum	VeriScore
SciNews	<code>Mistral_{FT}</code>	15.56	46.12	0.56
	<code>Lead-3_Q</code>	15.47	46.12	0.59
	<code>Lead-K_Q</code>	15.50	46.15	0.52
	<code>Tail-3_Q</code>	15.41	45.80	0.50
	<code>Tail-K_Q</code>	15.40	45.85	0.50
	<code>Random-3_Q</code>	15.35	45.88	0.48
	<code>Random-K_Q</code>	15.33	45.94	0.47
	<code>All-EDUs_Q</code>	15.48	46.11	0.52
	<code>NonExp-EDUs_Q</code>	15.42	46.03	0.49
	<code>Mistral_{ZS+PG}</code>	7.15	37.11	0.46
eLife	<code>Mistral_{FT}</code>	14.40	47.68	0.58
	<code>Lead-3_Q</code>	14.68	48.14	0.59
	<code>Lead-K_Q</code>	14.71	48.11	0.60
	<code>Tail-3_Q</code>	14.25	47.85	0.54
	<code>Tail-K_Q</code>	14.30	47.74	0.54
	<code>Random-3_Q</code>	14.65	48.05	0.56
	<code>Random-K_Q</code>	14.62	48.00	0.55
	<code>All-EDUs_Q</code>	14.62	48.10	0.56
	<code>NonExp-EDUs_Q</code>	14.56	48.02	0.52
	<code>Mistral_{ZS+PG}</code>	10.11	38.69	0.42
PLOS	<code>Mistral_{FT}</code>	14.31	40.22	0.57
	<code>Lead-3_Q</code>	14.69	40.66	0.59
	<code>Lead-K_Q</code>	14.57	40.59	0.56
	<code>Tail-3_Q</code>	13.78	40.12	0.52
	<code>Tail-K_Q</code>	13.72	40.18	0.51
	<code>Random-3_Q</code>	14.21	40.26	0.54
	<code>Random-K_Q</code>	14.14	40.20	0.53
	<code>All-EDUs_Q</code>	14.41	40.33	0.55
	<code>NonExp-EDUs_Q</code>	14.37	40.33	0.54
	<code>Mistral_{ZS+PG}</code>	6.74	32.44	0.48

Table 6.3: Comparison of different plan generation strategies using Plan-Input and `Mistral` in fine-tuned and zero-shot settings.

Impact of Discourse Parser. We next investigate the extent to which the choice of RST parser influences summary output quality. Specifically, we compare the DMRST

parser, which has been trained on the RST treebank, against results derived from GPT-4o and Mistral in zero-shot settings (the instruction is detailed in Appendix Figure D.6). For GPT-4o and Mistral, we use these models to directly extract explanatory sentences and their target sentences, rather than performing RST parsing, as RST parsing is a more complex task.⁴ Following Stede et al. (2017), we implement a rule-based method (RB) for extracting explanations (detailed in Appendix Section D.1). We also use the RST-Coref parser (Guz et al. 2020) and LLaMA-based RST parser (Maekawa et al. 2024) for comparison. To simulate parser instability, we design a random replacement (RR) method, where explanatory EDUs identified by DMRST are randomly replaced with non-explanatory or non-target EDUs. The number of replacements varies from one to the total number of explanatory EDUs per article, mimicking parser inaccuracy in identifying explanatory content. Finally, we introduce full random replacement (FRR), a variant representing the worst-case scenario of parser inaccuracy, where all explanatory EDUs are replaced with randomly selected non-explanatory or non-target EDUs. All these models are trained using Plan-Input as the default model.

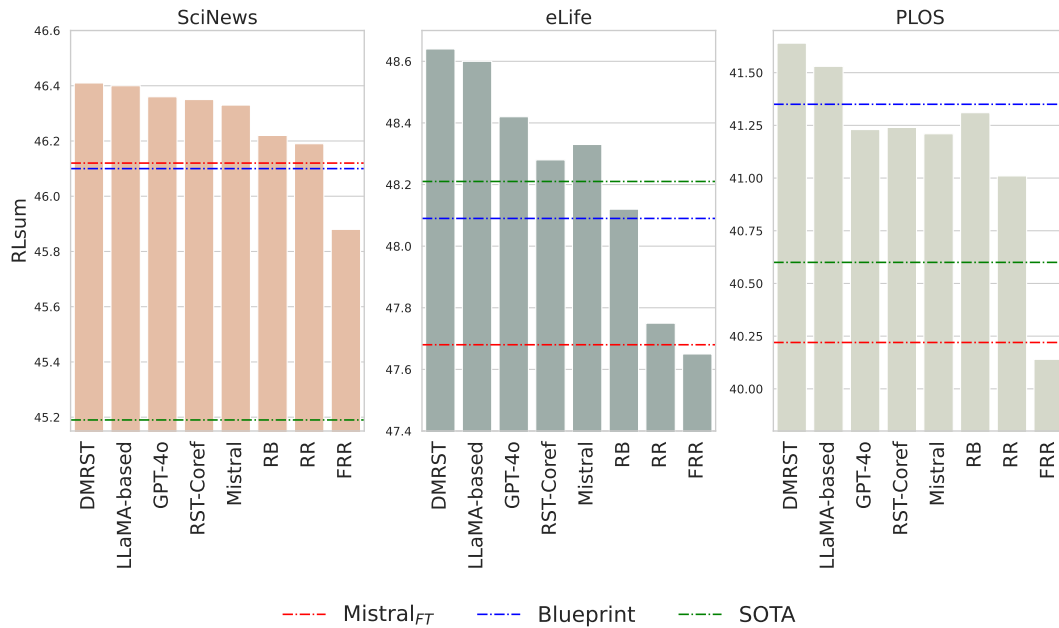


Figure 6.3: Summary quality as a function of different RST parsers.

Figure 6.3 shows that summarization performance improves with parsing accuracy (see DMRST vs. RR/FRR). DMRST consistently outperforms more naive methods based

⁴ The performance is better in the former way—on average, the RLsum score is 7.82% higher than the latter across datasets when applied Mistral under zero-shot setting.

on GPT-4o and Mistral, as well as the RST-Coref parser and the rule-based approach. Its performance is comparable to the LLaMA-based RST parser, but it has a smaller parameter size and incurs lower computational costs in parsing. We also find that RST-based models (irrespective of the parser at hand) outperform the baseline fine-tuned (Mistral_{FT}) and SOTA models (the performance of the previous best-published model on RLsum for each dataset in Table 6.2), and frequently surpass the Blueprint_{MT} model, as illustrated by the red, green, and blue dashed lines, respectively.

Impact of Plan Quality. To evaluate how the quality of plan questions affects model performance, we employ GPT-4o in a zero-shot setting as a question generator. In addition, we apply the Mistral model and the recent question generation model (RAST), developed by Gou et al. (2023). Moreover, we implement another random replacement (RR) method, where questions generated by GPT-4o are randomly substituted with irrelevant ones.⁵ The number of questions replaced varies from one to all for each article. Finally, we introduce another full random replacement (FRR), where questions generated by GPT-4o are all replaced with random irrelevant questions. Throughout these experiments, Plan-Input remains the default model.

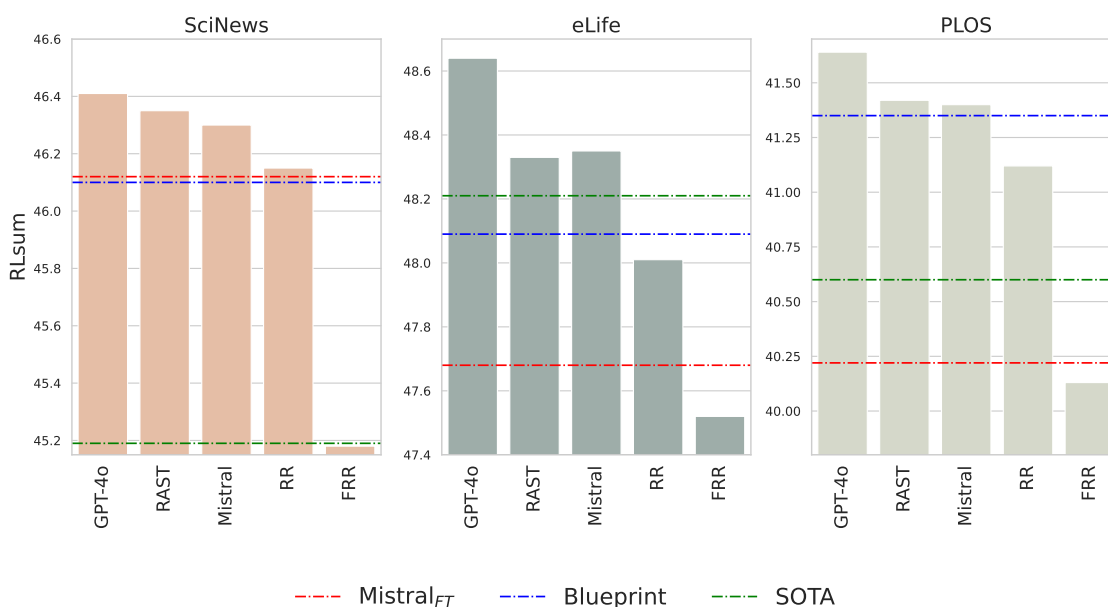


Figure 6.4: Summary quality as a function of different question generation methods.

⁵ The prompt can be found in Appendix Figure D.7.

Figure 6.4 reveals that plan questions have a direct impact on summary quality (compare GPT-4o vs. RR/FRR). Our method exhibits a certain degree of robustness, as it performs reasonably well even when the questions contain some degree of noise (RR vs. FRR). In the FRR condition, model performance degrades substantially, which suggests that relevant plan questions are instrumental in improving summary quality. Although questions generated by Mistral and RAST lead to lower performance gains compared to GPT-4o, these open-source models offer a cost-effective alternative. Despite not reaching the highest performance, they still surpass the previous state-of-the-art results.

Impact of Plan Composition. An advantage of plan-based models is their capacity to directly control the content of the output summary by simply altering the plan that gives rise to it. In this section, we investigate the extent to which changes in the formation of the plan questions can be observed in the generated summaries.⁶ To this end, we conduct a controlled generation experiment. We investigate whether our models can be guided to avoid generating these particular types of explanations by eliminating questions from the plan that correspond to specific types of explanatory EDUs.

Specifically, we remove from the plan questions corresponding to *Background*, *Comparison*, *Elaboration*, and *Explanation* (EDUs are identified by the DMRST parser). Table 6.4 reports the proportion of explanatory EDUs in the output summaries for our two models when no manipulation takes place. These results are compared with summaries generated from plans in which specific types of explanations have been deliberately removed. By deleting plan questions related to certain types of explanations, we evaluate the models' capacity to refrain from generating particular types of EDUs. Results presented in Table 6.4 confirm that our discourse-driven planning approach is sufficiently expressive to control the explanations and their discourse function in the summary. For both Plan-Input and Plan-Output, we observe a close alignment between the questions in the plan and the types of explanations produced in the summary.

⁶ We modify the plan generated by the model and evaluate whether our modifications have any bearing on the output.

Data	Relation	Plan-Output			Plan-Input		
		NoDel	Del	Δ RLSum	NoDel	Del	Δ RLSum
SciNews	Background	4.04	0.98	0.45	4.18	1.07	0.42
	Comparison	2.21	0.21	0.38	2.60	0.33	0.39
	Elaboration	6.47	1.25	1.56	6.39	1.20	1.50
	Explanation	7.09	1.69	1.62	7.88	2.11	1.56
eLife	Background	4.38	1.04	0.27	4.24	1.02	0.24
	Comparison	2.40	0.17	0.13	2.59	0.19	0.15
	Elaboration	5.46	1.16	0.46	5.42	1.15	0.40
	Explanation	6.07	1.72	0.58	6.11	1.81	0.52
PLOS	Background	4.33	0.97	0.20	4.09	0.92	0.22
	Comparison	2.02	0.09	0.11	2.12	0.13	0.13
	Elaboration	5.25	1.13	0.26	5.32	1.15	0.25
	Explanation	6.42	1.52	0.48	6.72	1.49	0.42

Table 6.4: Proportion of explanatory EDUs in predicted summaries for models with (Del) and without (NoDel) deletions in the plan. In the NoDel setting, the model regulates the proportion of explanatory content in the plan and summary. In the Del setting, the plan is manipulated by removing specific explanations. Δ RLSum = RLsum before plan deletion – RLsum after plan deletion.

Amongst explanatory relations, *Elaboration* and *Explanation* emerge as the most significant contributors to overall summary quality. In the Plan-Input model, the deletion of *Elaboration* explanations consistently results in lower RLsum across datasets (e.g., a drop from 46.41 to 44.91 in SciNews and from 48.64 to 48.24 in eLife). In contrast, removing *Comparison* explanations has a minor effect on summary quality, as reflected in the relatively stable RLsum scores across datasets (e.g., 41.64 vs. 41.51 in PLOS and 48.64 vs. 48.49 in eLife). In the NoDel condition, we observe that *Elaboration* and *Explanation* relations are more prominent in the summary texts, which suggests that removing them will be more detrimental (e.g., compared to *Background* or *Comparison*).

Human Evaluation. To alleviate the limitations of automated metrics and more profoundly analyze the quality of model-generated summaries, we randomly select 5 instances from each of the SciNews, eLife, and PLOS datasets (a total of 15) for human evaluation. The evaluators are postgraduate or doctoral students specializing in Computer Science or Computational Linguistics with advanced proficiency in English. They are compensated at our university’s standard hourly rate and remain blind to the origin of each summary during the evaluation process. In this study, we compare Mistral_{FT},

Blueprint_{MT}, Plan-Input (our best performing model), GPT-4o_{ZS}, and human-authored reference summaries. Each summary is independently rated by three different evaluators, resulting in a total of 225 ($15 \times 5 \times 3$) evaluation samples.

Our judges rate each summary on a 1 to 5 Likert scale across the dimensions of Faithfulness, Relevance, Informativeness, Accessibility, Explanation Accuracy, and Explanation Usefulness, with higher scores indicating better quality. They are also tasked with ranking the summaries from highest to lowest based on their relative quality within the set of candidates. Raters highlight explanations while reading the summary and the source document. To assess the explanations’ accuracy and usefulness, judges are allowed to consult external knowledge sources, such as books, academic papers, and Wikipedia, but not any AI tools. When they are not able to judge the accuracy or usefulness of explanations, raters are asked to skip them and flag them for expert review. We provide detailed evaluation guidelines in Appendix Section D.4 and discuss cases flagged for expert review in Appendix Section D.2.

Figure 6.5a reports the performance of each model across all summary samples, as well as the proportion of times a model is rated best (or worst). Fleiss’ Kappa scores for Faithfulness ($\kappa = 0.712$), Relevance ($\kappa = 0.835$), Informativeness ($\kappa = 0.673$), Accessibility ($\kappa = 0.707$), Explanation Accuracy ($\kappa = 0.604$), and Explanation Usefulness ($\kappa = 0.633$) indicate substantial inter-rater agreement, with an average of $\kappa = 0.694$. Overall, we observe that human-written summaries outperform all neural summarization models in terms of quality, with particularly pronounced differences in the usefulness and accuracy of explanations. Notably, human-written summaries also perform best in terms of faithfulness.

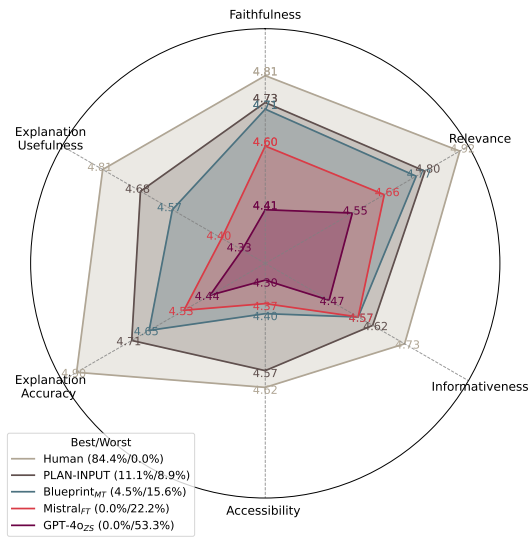
Among the four neural models under consideration, GPT-4o_{ZS} performs worst, while Mistral_{FT} also obtains low results, with a 22.2% likelihood of being rated as worst. Models based on planning (Blueprint_{MT} and Plan-Input) outperform Mistral_{FT}, with Plan-Input being superior across all metrics. On certain criteria, such as accessibility, Plan-Input is on par with human summaries. Compared to other neural summarization systems, it is also more likely to generate high-quality summaries.

LLM-as-a-Judge Evaluation. Due to the considerable length of the source documents and their summaries, it is not feasible to conduct extensive manual evaluations. Therefore, we also use an LLM-based evaluator (Liusie et al. 2024; Zheng et al. 2023; Liu et al. 2025a) to perform large-scale comparisons of system outputs. We use the

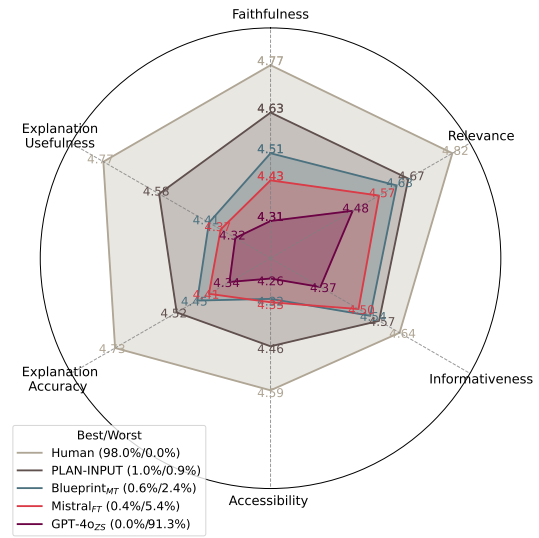
prompt targeting the same dimensions of summary quality adopted in our human evaluation and use GPT-4o as our evaluator. To avoid potential bias from previous interactions, we reset the conversation history before each query, making no further adjustments.

Firstly, we validate the agreement between GPT-4o and human ratings using the same set of 15 samples originally employed for human evaluation. We obtain a single rating per sample by averaging the scores of individual participants. We compute Fleiss' Kappa to measure the agreement between GPT-4o scores and the aggregated human ratings across evaluation dimensions. The resulting agreement scores are as follows: Faithfulness ($\kappa = 0.582$), Relevance ($\kappa = 0.643$), Informativeness ($\kappa = 0.633$), Accessibility ($\kappa = 0.624$), Explanation Usefulness ($\kappa = 0.615$), and Explanation Accuracy ($\kappa = 0.597$). We find that human raters and GPT-4o are in substantial agreement. Following this, we expand the evaluation to include all samples in the test sets across our three datasets.

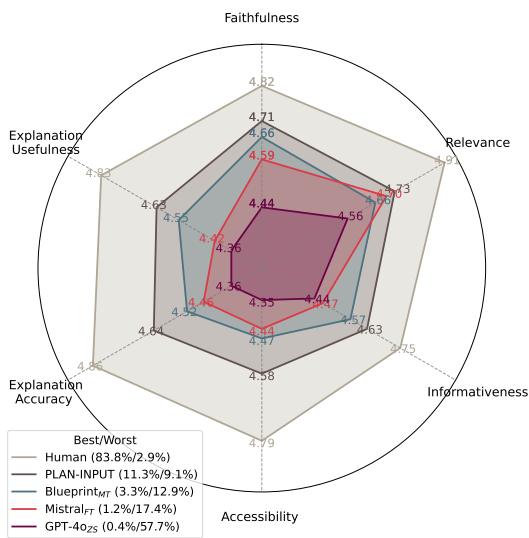
We present results on SciNews in [Figure 6.5b](#) and in [Figure 6.5c](#) and [Figure 6.5d](#) for the other two datasets. GPT-4o assigns the lowest scores to its own answers compared to fine-tuned models. Human-written summaries consistently receive the highest rating and are generally regarded as of the highest quality. In line with our human evaluation, GPT-4o also recognizes that Plan-Input is better than Blueprint_{MT} and Mistral_{FT}.



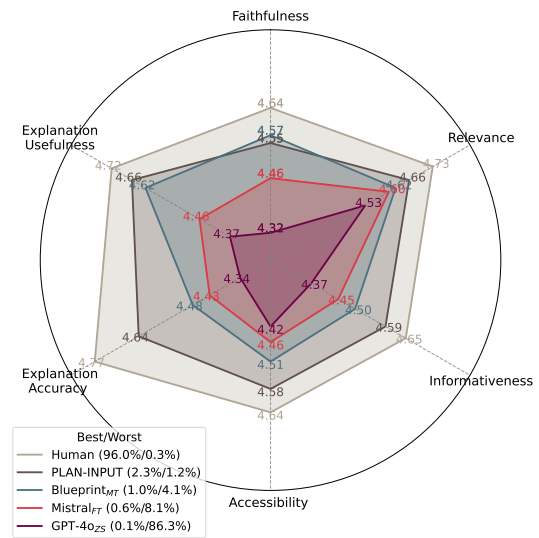
(a) Human evaluation results along different dimensions of summary quality.



(b) Evaluation with GPT-4o on SciNews.



(c) Evaluation on eLife using GPT-4o as a judge.



(d) Evaluation on PLOS using GPT-4o as a judge.

Figure 6.5: Human and GPT-4o-based evaluation results across datasets and evaluation dimensions.

Significance Testing. We use paired t-tests to analyze the results of our human evaluation and automatic evaluation (LLM-as-a-Judge with GPT-4o). In the human evaluation, we find that human responses are perceived as significantly better against all baseline models across metrics; with regard to Faithfulness and Accessibility, differences against Plan-Input and Blueprint_{MT} are not statistically significant. Plan-Input is significantly better in Explanation Usefulness and Accessibility compared to all neural network models (but not human answers).

In the LLM-as-a-Judge evaluation, human summaries outperform all neural models across all datasets and metrics with statistical significance. Plan-Input is significantly better over other neural network models on the SciNews dataset across all metrics except Relevance and Informativeness. On the eLife dataset, Plan-Input outperforms other neural network models on all metrics except Faithfulness, Explanation Usefulness, and Relevance. On the PLOS dataset, Plan-Input significantly outperforms competitor models in all metrics except Informativeness, Faithfulness, and Relevance. These results are derived with a significance level of $p < 0.05$.

Qualitative Analysis. We conduct a qualitative analysis by examining two case studies (see Table 6.5 and Table 6.6), with a particular focus on comparing outputs generated by humans and machines. Through this analysis, we observe differences in hallucination patterns across different models, and our findings suggest that incorporating structured planning into the generation process can effectively reduce certain types of errors.

Specifically, plan-based models demonstrate improved factual consistency compared to GPT-4o and Mistral_{FT} by better organizing information before generation. For instance, when describing how the cerebellum adjusts movements based on feedback (see Table 6.5), GPT-4o states, “The cerebellum acts as the brain’s command center for motor control.” At first glance, this phrasing may seem reasonable, but it inaccurately represents the cerebellum’s role and could lead to misconceptions. Instead of commanding movements, the cerebellum fine-tunes them using feedback from previous actions, adjusting their timing and coordination. When summarizing a document focusing on Musashi-1, a protein that binds to molecules of RNA and helps to promote cell growth during development (see Table 6.6), Mistral_{FT} incorrectly attributes certain properties to Musashi-1, such as it being “permanently destroyed Musashi-1” by oleic acid. In reality, Musashi-1 is not permanently destroyed but rather inhibited or down-regulated by compounds like oleic acid. These mistakes suggest a tendency to overgeneralize pat-

terns or infer causal relations that are not supported by evidence. In contrast, plan-based models (Plan-Output and Plan-Input) avoid these types of errors.

Unlike GPT-4o and Mistral_{FT}, which often produce fragmented or overly general statements, our models demonstrate a better grasp of causal and logical connections within the scientific narrative. For example, in the cerebellum case (see Table 6.5), our models successfully preserve the logical flow between experimental observations and their implications, while GPT-4o distorts causal relations by overgeneralizing. This is because structured planning forces the model to first outline key arguments and causal relations before generating text. This process reduces the risk of introducing spurious causal or conceptual connections, which are particularly important for explanation-oriented text.

Plan-based models are not without limitations; for instance, we occasionally observe misattributions, such as linking Musashi-1 to kidney cancer (see Table 6.5). However, they offer an improvement over models like GPT-4o in both factual accuracy and overall reliability. In sum, the qualitative analysis underscores the potential of plan-based models to enhance lay summarization quality while also highlighting areas for further refinement.

Model	Output
Human Reference	The cerebellum utilizes proprioceptive feedback to fine-tune the timing of movements in a sequence based on previous actions. Imagine the cerebellum as a coach who watches how you perform a move, then gives tips to improve the next one based on what was seen. But how exactly does it achieve this? To investigate, we trained rabbits to blink in response to an external cue and explored whether the cerebellum could use feedback from one blink to trigger the next. As expected, after learning the initial blink, the rabbits blinked again in response to their own first blink, creating a chain of movements. Control experiments confirmed that each blink was initiated by the previous one rather than the original cue. Consistent patterns of brain activity during this process indicate that the cerebellum adjusts movement based on feedback from previous actions. Building on this, we trained rabbits to blink on cue, and they learned to initiate additional blinks in response to earlier blinks in the sequence. We further found that the rabbits could use a blink from one eye as a cue to trigger a blink in the other eye, suggesting that the same mechanism governs these movements. This raises the possibility that the cerebellum might also guide sequences of cortical activity during cognitive tasks, given its extensive connections to the cortex, a question future experiments should explore.
GPT-4o	The cerebellum acts as the brain's command center for motor control , directly issuing movement instructions and ensuring precise timing. To test this, scientists trained rabbits to blink in response to a vibration signal . Once trained, the rabbits began blinking in a continuous rhythm, even without an external cue, indicating that the cerebellum could independently generate movement sequences. Interestingly, when the researchers covered one of the rabbit's eyes, the uncovered eye still continued blinking, suggesting that the cerebellum maintains a direct, hardwired link between the two eyes. These findings demonstrate that the cerebellum is not only responsible for movement execution but may also function as a memory hub for motor programs, completely independent of the cerebral cortex.
Mistral _{FT}	The cerebellum helps coordinate movements by adjusting actions based on prior attempts. To study this, scientists trained rabbits to blink in response to a sound. Over time, the rabbits developed a habit of blinking twice in succession, suggesting that the cerebellum is responsible for generating repetitive movements . Further experiments showed that rabbits could use one blink as a cue for another, even switching between eyes. This suggests that the cerebellum not only controls basic motor sequences but also plays a role in synchronizing different body movements . It is even possible that the cerebellum helps regulate general cognitive functions , though more research is needed.
Plan-Output	<p>q1: What can the role of the cerebellum in motor control be compared to? q2: How do rabbits form a sequence of consecutive blinks after training? q3: What does it suggest that rabbits can use a blink from one eye to trigger a blink in the other eye?</p> <p>The cerebellum acts like a coach, refining movements based on past experiences. To explore this, scientists trained rabbits to blink when they heard a tone. Over time, the rabbits learned to initiate a second blink in response to their own first blink, forming a sequence of movements. This suggests that the cerebellum uses sensory feedback to fine-tune motor control. Furthermore, the rabbits could blink one eye and then trigger a blink in the other eye, indicating that the cerebellum directly controls bilateral facial coordination. These findings highlight the cerebellum's essential role in sequencing motor actions and suggest that it may also contribute to higher-level cognitive processes.</p>
Plan-Input	<p>q1: What kind of role does the cerebellum play in motor control? q2: How do rabbits trigger a sequence of blinks from an initial blink after training? q3: What does it reveal about motor control mechanisms that rabbits can use a blink from one eye to trigger a blink in the other eye?</p> <p>The cerebellum functions as a predictive system that learns movement patterns and automatically sequences them. Researchers trained rabbits to blink when given an air puff. After sufficient training, the rabbits began to produce consecutive blinks, with each blink triggering the next. This suggests that the cerebellum can store movement chains and replay them when needed. Additionally, the researchers found that a blink in one eye could lead to a blink in the opposite eye, implying that the cerebellum might have specialized circuits dedicated to synchronizing both sides of the body. These insights open the possibility that the cerebellum also orchestrates sequences of thoughts and decisions, a topic for future research.</p>

Table 6.5: Case study 1: we use red font to mark hallucinated information. This sample comes from the test split of the eLife dataset (Goldsack et al. 2022).

Model	Output
Human Reference	<p>When an embryo is developing, stem cells must divide and develop into many specialized types of cells. However, if cell division doesn't stop, or if it restarts later in life, it can cause tumors to form. Musashi-1 is a protein that binds to molecules of RNA and helps to promote cell growth during development: mice that lack this protein have serious brain defects and die shortly after birth. Musashi-1 is usually turned off in adult cells that are not dividing. Sometimes, however, it remains active and contributes to the growth of cancers in the brain and the gut. Reducing Musashi-1 levels in colon tumors slows their growth and causes the cancer cells to die. To find a compound that would switch off Musashi-1, Clingman et al. screened more than 30,000 compounds and identified four inhibitors. One of these was oleic acid, a fatty acid that is found in olive oil and other animal and plant oils. Oleic acid interacts with Musashi-1 in a way that changes the shape of the protein. These changes mean that Musashi-1 is no longer able to regulate the genes that control cell proliferation. Clingman et al. also found that Musashi-1 promotes the activity of a particular enzyme that makes fatty acids; molecules that are needed in large quantities when cells are dividing. Musashi-1 appears to act as a nutrient sensor, turning down the activity of this enzyme in cells when levels of oleic acid are high, and turning up enzyme activity when oleic acid levels are low. The findings of Clingman et al. further reveal how our diets can affect gene expression, and may aid the development of new treatments against cancer.</p>
GPT-4o	<p>Musashi-1 is a protein that plays a crucial role in early development by directly binding to DNA and activating genes responsible for cell proliferation. In mice that lack Musashi-1, the protein is compensated by other molecular pathways, allowing them to develop normally with only minor cognitive impairments. To explore ways to inhibit Musashi-1, Clingman et al. analyzed a vast library of nearly 135,000 chemical compounds. They identified four potential inhibitors, including oleic acid, a fatty acid mainly found in red meat and dairy products. This compound works by permanently destroying Musashi-1, preventing it from affecting cell division. In cancer studies, researchers found that reducing Musashi-1 levels in colon tumors leads to uncontrolled tumor growth and increased metastasis. These findings suggest that Musashi-1 may function as a critical tumor suppressor, and future therapies should aim to enhance its activity in cancer patients rather than inhibit it.</p>
Mistral _{FT}	<p>Musashi-1 is a protein that regulates cell division by binding to RNA. It is active during embryonic development but usually turns off in adults. However, it can remain active and contribute to tumors in the brain and lungs. To inhibit Musashi-1, researchers screened over 30,000 compounds and found oleic acid, a substance present in coconut oil and dairy products. Oleic acid interacts with Musashi-1 and permanently destroys it, stopping it from regulating genes linked to cell growth. Additionally, high oleic acid levels completely shut down metabolism in cells, showing how diet influences gene regulation and cancer treatment.</p>
Plan-Output	<p>q1: What is the role of Musashi-1 in cell development and cancer growth? q2: How does oleic acid affect Musashi-1 and its function in cell proliferation? q3: How does Musashi-1 function as a nutrient sensor in relation to fatty acid synthesis?</p> <hr/> <p>Musashi-1 is a protein that binds to RNA molecules and plays a crucial role in cell division and tissue development. During early embryonic growth, it ensures proper differentiation of stem cells. Mice lacking Musashi-1 experience severe neurological defects and typically survive for several months before succumbing to developmental issues. Normally, Musashi-1 is turned off in mature cells, but in some cases, it remains active and has been linked to tumor formation in the brain and kidneys. To counteract Musashi-1 activity, Clingman et al. screened over 30,000 potential inhibitors and identified four promising compounds. One of them, oleic acid—found in olive oil and fish oil—was particularly effective. It interacts with Musashi-1 in a way that causes the protein to break down, thereby preventing it from controlling genes involved in cell proliferation. Additionally, Musashi-1 functions as a metabolic sensor by regulating fatty acid synthesis. When oleic acid levels are low, Musashi-1 suppresses enzyme activity, reducing fatty acid production. However, when oleic acid concentrations rise, the enzyme becomes highly active, increasing fatty acid availability to support rapid cell growth. These findings illustrate the link between diet and gene regulation and could inform new therapeutic strategies for cancer treatment.</p>
Plan-Input	<p>q1: What happens to mice that do not have Musashi-1, and what does this indicate about its function? q2: How did researchers identify inhibitors of Musashi-1, and what is one of the inhibitors they found? q3: What effect does reducing Musashi-1 levels have on colon tumors, and what does this suggest for cancer treatment?</p> <hr/> <p>Musashi-1 is a crucial protein that binds to RNA and plays a key role in cell division and development. When mice lack Musashi-1, they develop mild brain defects but survive into adulthood, though with noticeable neurological impairments. This suggests that Musashi-1 is essential for normal brain development and proper neural function. To identify potential inhibitors of Musashi-1, Clingman et al. conducted a large-scale screening of nearly 52,000 chemical compounds. Among the four inhibitors they discovered, one was oleic acid, a fatty acid commonly found in olive oil and coconut oil. Oleic acid interacts with Musashi-1 by binding to it directly, leading to its complete deactivation. Reducing Musashi-1 levels in colon tumors has been shown to accelerate their growth, making the tumors more aggressive. This suggests that Musashi-1 may have a dual role in cell regulation, potentially acting as both a tumor promoter and suppressor in different contexts. These findings provide new insights into the complex role of Musashi-1 in cancer and could inform the development of novel therapeutic strategies.</p>

Table 6.6: Case study 2: we use red font to mark hallucinated information. This sample comes from the test split of the eLife dataset (Goldsack et al. 2022).

6.5 Related Work

Lay Summarization. Lay Summarization involves transforming complicated scientific texts into accessible summaries for non-experts, thereby enhancing public understanding of scientific research. Existing investigations have tackled several challenges, such as creating datasets and developing task-specific models. Chandrasekaran et al. (2020) pioneered the task of scientific lay summarization, while follow-up work focused on the development of larger datasets, including CDSR (Guo et al. 2021b), eLife, PLOS (Goldsack et al. 2022), and SciNews (Liu et al. 2024b). Methodological advancements include integrating domain-specific knowledge graphs with encoder-decoder models (Goldsack et al. 2023) and developing methods like ATLAS (Zhang et al. 2024c), which facilitate the generation of summaries with varying degrees of accessibility for lay audiences. However, none of the existing approaches emphasize discourse-driven mechanisms for modulating summary content, nor do they focus on evaluating the quality or proportion of explanatory content.

Interactive Summarization. Interactive Summarization refers to the process of enabling users to engage with and influence the summarization process through interactive interfaces (Costa et al. 2018; Guo et al. 2021a; August et al. 2023; Kim et al. 2019; Fok et al. 2024). For example, August et al. (2023) introduced an interface that provides localized plain-language key points, definitions, and question-based navigation, allowing users to explore medical research more effectively. Similarly, Fok et al. (2024) proposed a design for recursively expandable abstracts, where users can click on keywords or phrases to access system-generated text and citation links. While interactive summarization is not the focus of our work, it is worth pointing out that users could choose to modify the plan and influence the generation process.

RST in Summarization. Rhetorical Structure Theory (RST) is a discourse framework that analyzes the relations between text units, termed Elementary Discourse Units (EDUs), and their roles within the document (Marcu. 1997; Marcu. 1999b; Marcu. 2000). RST categorizes connections among EDUs into different rhetorical or coherence relations, such as *Elaboration*, *Contrast*, and *Causality*. Within this framework, EDUs are arranged into a hierarchical tree structure, where nucleus EDUs are supported or elaborated by satellite EDUs. Early research (Marcu. 1997; Marcu. 1999b; Louis

et al. 2010) has found that human-written summaries are often derived from nucleus EDUs, affirming RST’s usefulness for summarization tasks. Several studies have explored ways to inject discourse structure into summarization models and demonstrated improved performance compared to models without such discourse injection (Xu et al. 2020; Dong et al. 2021; Chen et al. 2021; Liu et al. 2023c; Liu et al. 2024a). Our work applies RST to discern explanatory EDUs, which elucidate or support propositions, and their related target (explained) EDUs within reference summaries.

Planning in Summarization. Text summarization with planning involves organizing content through structured representations prior to generating summaries. Existing work has primarily focused on *phrase-level* planning, which has been shown to improve summary quality and reduce hallucination (Narayan et al. 2021; Narayan et al. 2023). Plans have previously taken the form of named entities (Narayan et al. 2021; Liu et al. 2021d; Huot et al. 2024), keyword prompts (Creo et al. 2023), and question-answer pairs, referred to as blueprint (Narayan et al. 2023). Our work introduces *EDU-level* planning, placing emphasis on larger semantic units. Our plans take the form of questions (without answers), triggered by explanations identified through RST. This definition is more flexible (compared to question-answer pairs), as we do not require answers to be incorporated and are not limited to a specific style of answers or questions.

Question Under Discussion. Question Under Discussion (QUD) conceptualizes discourse as a dynamic interaction of continuously posed and answered questions (Roberts. 2012; Benz et al. 2017; De Kuthy et al. 2020; Ko et al. 2023). It can be seen as a question generation strategy, where each sentence serves as a response to an implicit question arising from its preceding context, thereby delineating the text’s intent and structure (Beaver et al. 2017; Wu et al. 2023a; Wu et al. 2023b). Although QUD has been explored in discourse analysis (Ko et al. 2023; Wu et al. 2023a) and to interpret complex linguistic phenomena like presupposition and information structure (Beaver et al. 2017), its application in text generation and summarization has received limited attention. Wu et al. (2023b) leveraged manual annotation to recover implicit questions to guide sentence-level simplification. Our work draws inspiration from QUD in formulating plans as a series of questions. We work with more complex documents and rely on large language models, bypassing issues of scalability arising from manual annotations.

6.6 Chapter Summary

This chapter addressed Research Question 4 by investigating whether explanatory content in lay summarization can be treated as a controllable discourse element whose placement, scope, and function are explicitly managed through planning. Building on the QUD-based planning framework introduced in the preceding chapter, this chapter combined RST-based identification of explanatory EDUs with question-driven plans that target specifically the explanatory portions of lay summaries. Two model variants were developed to test different integration strategies, with the plan serving either as an input condition or as a generation prefix. Experiments on SciNews, eLife, and PLOS showed that EDU-level discourse planning tends to produce summaries with a higher proportion of explanatory content, improved readability, and stronger factual consistency compared to both vanilla fine-tuning and phrase-level planning baselines, providing empirical support for Hypothesis IV. Controlled generation experiments further demonstrated that modifying the plan, for instance by removing questions corresponding to specific relation types, leads to predictable changes in the generated explanations, confirming a degree of controllability that is difficult to achieve with end-to-end approaches.

With this chapter, the preceding four technical studies trace a trajectory from distributional discourse representations for attention modulation in long-document encoding, through parameter-efficient adaptation under constrained update budgets, to macro-level planning for multimodal summarization, and to fine-grained control over explanatory content in public-facing summaries. The methods explored so far, however, all operate under the assumption that the relevant evidence is already contained within a given source input, whether a long document, a video recording, or a scientific article to be summarized. In many knowledge-intensive scenarios, the evidence needed to produce a faithful response should first be retrieved from an external corpus, and the retrieved passages typically arrive as structurally disconnected chunks without explicit rhetorical relations among them. The following chapter investigates whether the discourse modeling principles developed throughout this dissertation can extend to the retrieval-augmented generation setting to improve factual coherence and answer quality.

7

Discourse-Aware Retrieval-Augmented Generation

Chapter 3 through Chapter 6 progressively develop a discourse-guided framework for text generation under increasingly complex conditions. Chapter 3 and Chapter 4 integrate uncertainty-aware RST distributions into the encoding process, demonstrating that distributional discourse signals can improve content selection and organization in long-document summarization under both full fine-tuning and parameter-efficient adaptation settings, which supports Research Question 1 (RQ1), Hypothesis I, Research Question 2 (RQ2), and Hypothesis II. Chapter 5 and Chapter 6 shift discourse modeling from the encoding stage to the generation stage by introducing QUD-based planning as an explicit intermediate variable, showing that structured plans can guide macro-level organization in multimodal and explanatory summarization, consistent with Research Question 3 (RQ3), Hypothesis III, Research Question 4 (RQ4), and Hypothesis IV. A shared assumption across these chapters is that all relevant evidence is already contained within a given source input, whether a long document, a video recording, or a scientific article. In many knowledge-intensive scenarios, however, this assumption does not hold. The evidence needed to produce a faithful response is typically first retrieved from an external corpus, and the retrieved passages arrive as structurally disconnected chunks without the discourse continuity that characterizes coherent documents. Standard retrieval-augmented generation (RAG) pipelines concatenate these chunks as flat context, offering no mechanism for the generator to reason about the rhetorical relations within or across them. Against this backdrop, Research Question 5 (RQ5) asks how discourse

structure can be leveraged to organize such disconnected retrieved evidence, so that retrieval-augmented generation produces more coherent and factually consistent outputs.

To this end, [Chapter 7](#) tests Hypothesis V that explicitly constructing intra-chunk discourse trees and inter-chunk rhetorical graphs over retrieved evidence, combined with discourse-driven planning, can improve the factual consistency, coherence, and answer quality of retrieval-augmented generation compared to standard flat-context approaches. Concretely, [Chapter 7](#) adapts the two discourse modeling paradigms developed in the preceding chapters, namely RST-based structural analysis for capturing local and global rhetorical relations and QUD-inspired planning for organizing the generation process, to the retrieval-augmented setting. The proposed framework, *Disco-RAG*, constructs local RST trees within each retrieved chunk, infers a directed rhetorical graph across chunks to model inter-chunk relations such as support, contrast, and elaboration, and synthesizes these structural signals through a planning module into an ordered blueprint that guides the final answer generation. Subsequent experiments on Loong (knowledge-intensive long-context QA), ASQA (ambiguous question answering), and SciNews (lay summarization) evaluate whether this discourse-aware evidence organization yields measurable improvements over standard RAG pipelines and prior state-of-the-art methods. [Chapter 7](#) also examines the individual contributions of each structural component through ablation studies and assesses the framework’s sensitivity to structural noise through controlled perturbation experiments.

7.1 Problem Formulation

The advent of large language models (LLMs; [Touvron et al. 2023](#); [Team 2025](#); [OpenAI 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. 2025b](#); [Luo et al. 2025](#)), and text generation ([Duong et al. 2025](#); [Bigelow et al. 2025](#); [Que et al. 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](#); [Ding et al. 2026](#); [Chen et al. 2026](#)). Retrieval-

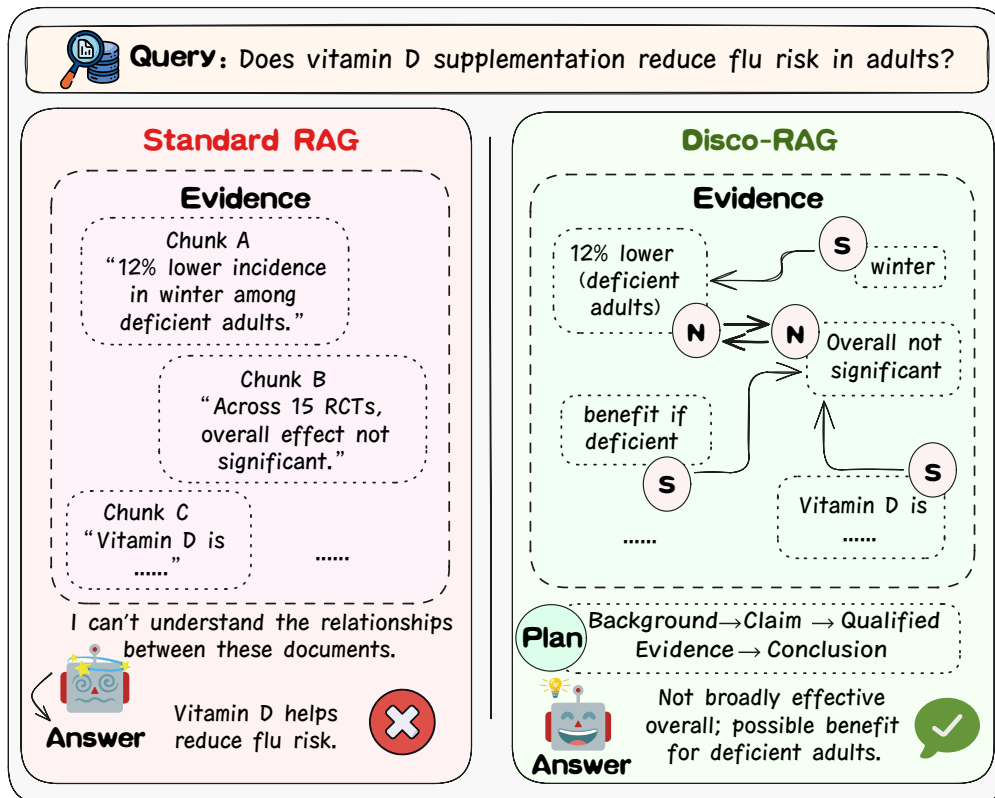


Figure 7.1: Comparison between standard RAG and Disco-RAG. While standard RAG retrieves isolated chunks without structural links, Disco-RAG organizes evidence into discourse structures (trees & graphs). Here, S denotes *Satellite* (the supplementary part), and N denotes *Nucleus* (the core part).

Augmented Generation (RAG) has been proposed as a suitable strategy by integrating an external knowledge component through retrieval-based mechanisms (Lewis et al. 2020b; Asai et al. 2024; Chan et al. 2024).

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. 2020b). 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. 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 iden-

tify rhetorical connections. As depicted in Figure 7.1 (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 et al. 2024a) 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 7.1 (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. 2024b). Consistent improvements are observed compared with standard RAG systems and state-of-the-art (SOTA) methods. On the Loong benchmark, our approach delivers an overall gain of 12.74 points in LLM Score. On the ASQA dataset, our method exceeds the best existing systems on Exact Match and ROUGE-L Score by clear margins. On the SciNews benchmark, Disco-RAG establishes new SOTA performance across most evaluation metrics.

7.2 Methodology

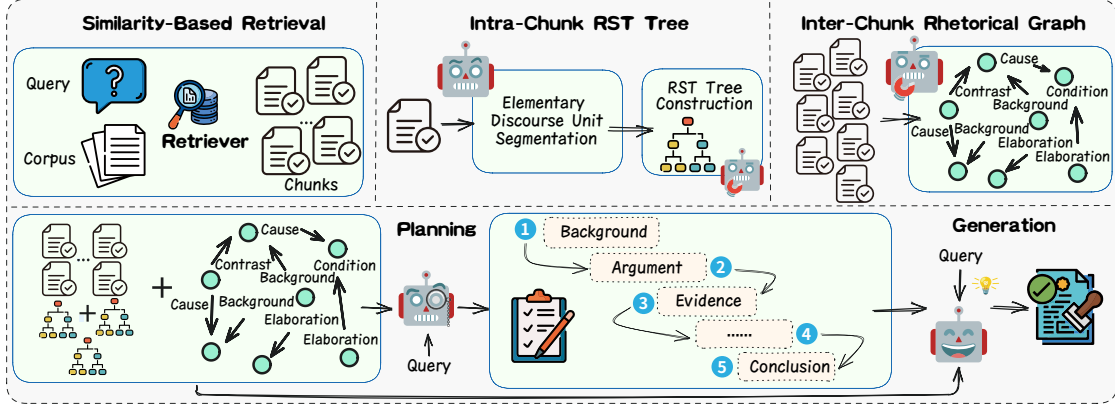


Figure 7.2: The Disco-RAG pipeline: Starting from passage retrieval (providing context), then intra-chunk RST tree parsing (capturing local discourse), inter-chunk rhetorical graph construction (modeling global discourse), rhetorical planning (blueprint generation), and answer generation (producing the final output).

Method Overview. We formalize the standard RAG as a conditional generation problem. Given a query q and a set of Top- k retrieved chunks $\mathcal{C}(q; \mathcal{D}) = \{c_1, c_2, \dots, c_k\}$ from a corpus \mathcal{D} , the output is $y = \arg \max_{y'} P(y' | q, \mathcal{C}(q; \mathcal{D}))$, where $P(\cdot)$ denotes the conditional distribution of the answer 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 7.2](#), 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.¹ Given c_i , parser \mathcal{A} jointly performs elementary discourse unit (EDU) segmentation and RST parsing, producing a sequence of EDUs $\{e_{i_1}, \dots, e_{i_m}\}$, nucleus and satellite role assignments, and rhetorical relations among EDUs. Formally, $c_i \xrightarrow{\mathcal{A}} t_i = (V_i, E_i)$, where $V_i = \{e_{i_1}, \dots, e_{i_m}\}$ 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 7.2](#) shows how EDUs are organized into hierarchical trees.²

The RST tree parsing is formalized as $P(t_i | c_i; \theta_{\mathcal{A}}) = \prod_{j=1}^m P(e_{i_j} | c_i; \theta_{\mathcal{A}}) \cdot \prod_{(u,v)} P(r_{u,v} | e_{i_u}, e_{i_v}; \theta_{\mathcal{A}})$, where $P(e_{i_j} | c_i)$ signifies the probability of EDU boundary prediction and $u, v \in V_i$ are discourse units, $P(r_{u,v} | e_{i_u}, e_{i_v})$ 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 $\mathcal{F} \subseteq \mathcal{C} \times \mathcal{C} \times (\mathcal{R} \cup \text{UNRELATED})$ 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 directed rhetorical relations $\{r_{i,j}\}$ or an UNRELATED label for all chunk pairs.³

The rhetorical graph construction is modeled as $P(\mathcal{G} | \mathcal{C}; \theta_{\mathcal{A}})$. This joint distribution can be factorized over ordered chunk pairs as $P(\mathcal{G} | \mathcal{C}; \theta_{\mathcal{A}}) = \prod_{i=1}^k \prod_{j=1, j \neq i}^k P(r_{i,j} | \mathcal{C}; \theta_{\mathcal{A}})$. As shown in the top-right panel of [Figure 7.2](#), 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 $(q, \mathcal{C}, \mathcal{T}, \mathcal{G}) \xrightarrow{\mathcal{A}} \mathcal{B}$.

¹ Prompt is detailed in Appendix [Figure E.3](#).

² Intra-chunk RST trees are constructed offline.

³ Appendix [Figure E.4](#) provides prompt and format details used in inter-chunk relation prediction.

As depicted in the center-bottom panel of [Figure 7.2](#), the plan \mathcal{B} is dynamically conditioned on the discourse structures and the query.⁴ 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 $y = \arg \max_{y'} P(y' | q, \mathcal{C}, \mathcal{T}, \mathcal{G}, \mathcal{B})$, where y' denotes a candidate output and y refers to the final output that maximizes the conditional probability.⁵

7.3 Experimental Setup

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. 2024b](#)). 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](#)).

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. 2024b](#)). For the Loong dataset ([Wang et al. 2024a](#); [Li et al. 2025c](#)), 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. 2020b](#)), SARI ([Xu et al. 2016](#)), and SummaC ([Laban et al.](#)

⁴ Appendix [Figure E.5](#) provides prompt used in discourse-aware planning.

⁵ Appendix [Figure E.11](#) contains the generation prompt.

2022). These metrics assess the informativeness, fluency, and factual consistency of generated answers. Detailed descriptions of these metrics are provided in [Section E.1](#).

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 ([Team. 2024](#)).⁶ 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 all compared methods.

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. 2020b](#))⁷, where relevant chunks are prepended to the query prior to inference.⁸ and (3) previously published results from state-of-the-art RAG (if applicable) systems on the same benchmark.

7.4 Results and Analysis

Main Results. The experimental results are summarized in [Table 7.1](#), [Table 7.2](#), and [Table 7.3](#), 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.

⁶ 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.

⁷ Prompts for full context generation and standard RAG method are provided in Appendix [Figure E.6](#) and [Figure E.7](#).

⁸ All experiments are training-free (unless otherwise specified) and use only task instructions. All hyperparameters follow the same settings as Disco-RAG.

Condition	Model	Spot.		Comp.		Clus.		Chain.		Overall	
		LLM Score _↑	EM _↑	LLM Score _↑	EM _↑	LLM Score _↑	EM _↑	LLM Score _↑	EM _↑	LLM Score _↑	EM _↑
<i>Set 1 (10K–50K Tokens)</i>											
<i>Full Context</i>	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
<i>Standard RAG</i>	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
<i>SOTA Results</i>	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. 2025c)	<u>74.53</u>	<u>0.47</u>	<u>75.58</u>	0.47	<u>65.13</u>	0.23	67.84	0.34	<u>69.43</u>	<u>0.35</u>
	Disco-RAG (Llama-3.1-8B)	73.35	0.40	73.57	0.37	64.44	0.12	68.01	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	69.39	0.33
	Disco-RAG (Llama-3.3-70B)	76.60	0.45	75.65	<u>0.45</u>	65.36	<u>0.17</u>	<u>68.30</u>	<u>0.38</u>	71.00	0.38
<i>Set 2 (50K–100K Tokens)</i>											
<i>Full Context</i>	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
<i>Standard RAG</i>	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
<i>SOTA Results</i>	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. 2025c)	<u>68.00</u>	0.41	63.71	0.36	<u>61.40</u>	<u>0.17</u>	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	<u>64.06</u>	<u>0.30</u>	60.63	0.15	<u>57.22</u>	<u>0.20</u>	<u>61.32</u>	<u>0.25</u>
	Disco-RAG (Llama-3.3-70B)	69.92	<u>0.39</u>	64.34	0.36	61.67	0.18	58.23	0.22	63.61	0.28
<i>Set 3 (100K–200K Tokens)</i>											
<i>Full Context</i>	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
<i>Standard RAG</i>	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
<i>SOTA Results</i>	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. 2025c)	68.62	0.44	<u>57.74</u>	0.35	<u>58.27</u>	0.10	49.73	0.13	<u>57.92</u>	<u>0.21</u>
	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	<u>51.20</u>	0.13	57.14	0.18
	Disco-RAG (Llama-3.3-70B)	<u>66.37</u>	<u>0.38</u>	57.84	<u>0.28</u>	58.85	<u>0.07</u>	52.17	<u>0.15</u>	58.86	0.22
<i>Set 4 (200K–250K Tokens)</i>											
<i>Full Context</i>	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
<i>Standard RAG</i>	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	<u>0.25</u>	34.49	0.02	36.41	0.01	31.33	0.02	35.61	0.07
<i>SOTA Results</i>	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. 2025c)	56.87	0.19	<u>55.62</u>	0.25	56.59	0.00	35.71	0.05	51.42	<u>0.10</u>
	Disco-RAG (Llama-3.1-8B)	56.68	0.19	53.92	0.12	57.53	<u>0.02</u>	36.00	0.03	50.87	0.08
	Disco-RAG (Qwen2.5-72B)	<u>57.27</u>	0.22	54.97	0.15	<u>57.40</u>	<u>0.02</u>	36.17	<u>0.06</u>	<u>54.47</u>	<u>0.10</u>
	Disco-RAG (Llama-3.3-70B)	57.74	0.27	55.81	<u>0.17</u>	57.36	0.03	<u>36.06</u>	<u>0.06</u>	54.62	0.11

Table 7.1: 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. (2025c). We use **bold red** to indicate the best results and blue underlined text to indicate the second-best results.

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 previously reported training-based method StructRAG (62.07 vs. 60.38).

Model	EM \uparrow	RL \uparrow	DR Score \uparrow
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. 2023b)	41.3	34.3	31.1
Tree of Clarifications (Kim et al. 2023)	—	39.7	36.6
Open-RAG (Islam et al. 2024b)	36.3	38.1	—
ConTReGen (Roy et al. 2024)	41.2	—	30.3
DualRAG (Cheng et al. 2025)	—	31.7	—
RAS (Jiang et al. 2026)	—	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	<u>42.2</u>	32.6
Disco-RAG (Qwen2.5-72B)	<u>41.8</u>	41.3	<u>33.2</u>
Disco-RAG (Llama-3.3-70B)	42.0	42.3	32.8

Table 7.2: Performance on the ASQA benchmark. Disco-RAG consistently outperforms standard RAG across all metrics. It also surpasses existing SOTA methods on most dimensions.

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, surpassing both standard RAG and the previous best system (Liu et al. 2024b; Liu et al. 2025b).

Model	RL \uparrow	BERTScore \uparrow	SARI \uparrow	SummaC \uparrow
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. 2024b)	20.12	62.80	41.56	—
SingleTurnPlan (Liang et al. 2024b)	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 7.3: Performance on the SciNews dataset. Disco-RAG beats both zero-shot and standard RAG, and often surpasses prior SOTA across multiple metrics.

Ablation Studies. We perform ablation studies on the Loong benchmark, as summarized in Table 7.4, 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 is removed, respectively. Similarly, the Ex-

Method	Set 1		Set 2		Set 3		Set 4		Overall	
	LLM Score \uparrow	EM \uparrow	LLM Score \uparrow	EM \uparrow	LLM Score \uparrow	EM \uparrow	LLM Score \uparrow	EM \uparrow	LLM Score \uparrow	EM \uparrow
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	0.18
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 7.4: 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.

act 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.⁹

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.

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

⁹ Prompts for these two generic planning baselines are provided in Appendix Figure E.8 and Figure E.9.

are visualized in Figure 7.3.

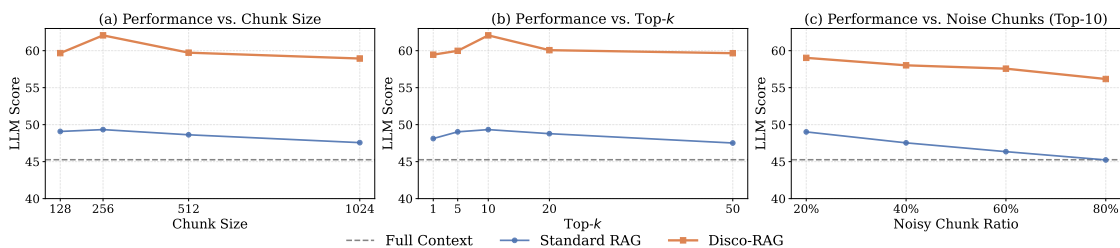


Figure 7.3: Performance comparison under varying chunk size (a), Top- k value (b), and retrieval noise level (c).

Panel (a) of Figure 7.3 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 58.91, showing strong robustness to granularity shifts. Panel (b) of Figure 7.3 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 7.3 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.

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.

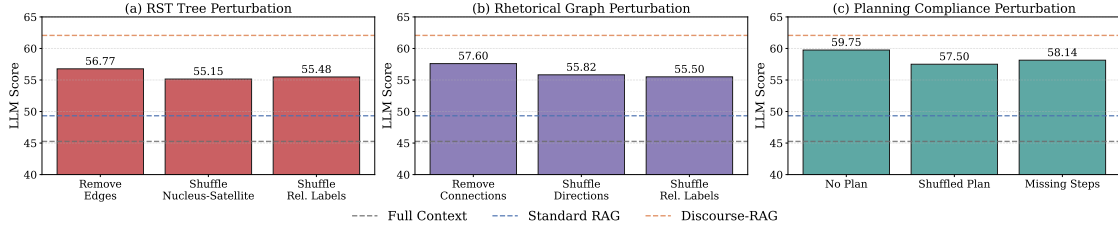


Figure 7.4: Effect of structural perturbations on performance. Panels (a), (b), and (c) correspond to intra-chunk RST trees, inter-chunk rhetorical graphs, and discourse-aware plans, respectively. Each perturbation involves randomly altering or removing the relevant elements.

Figure 7.4 presents the results of the perturbation study. Panel (a) of Figure 7.4 exhibits that perturbing intra-chunk structures leads to a 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 7.4 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 7.4 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 a 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.

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 7.5. Our zero-shot parser attains competitive scores that are close in nuclearity and relation prediction and somewhat lower in 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 7.5: 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 rhetorical 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 0.71 for intra-chunk discourse trees, 0.73 for inter-chunk rhetorical graphs, and 0.86 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.

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 a marker for each ordered chunk pair based on connective cues such as *however*, *but*, *although*, *in contrast*, *therefore*, *because*, *as a result*, and *meanwhile*.¹⁰

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

Table 7.6: Comparison of standard RAG, a shallow discourse marker variant, and Disco-RAG on the Loong benchmark with Llama-3.3-70B.

Table 7.6 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.

Mixed-Model Deployment. Since the structural modules (RST parsing, rhetorical graph construction, and planning) and the final generator are decoupled by design, they can be instantiated with different models. To examine whether a smaller model can serve

¹⁰The prompt used for shallow discourse marker inference is provided in Appendix Figure E.10.

as a cost-effective structural backbone, we conduct an experiment on the Loong benchmark in which Llama-3.1-8B handles all structural modules while Llama-3.3-70B is used only for the final generation stage. Table 7.7 reports the results under four configurations.

Configuration	Structurer / Generator	LLM Score _↑	EM _↑
Standard RAG	70B / 70B	49.33	0.17
Disco-RAG (all 8B)	8B / 8B	58.94	0.20
Disco-RAG (8B+70B)	8B / 70B	60.52	0.22
Disco-RAG (all 70B)	70B / 70B	62.07	0.24

Table 7.7: Mixed-model results on the Loong benchmark. *Structurer* refers to the model used for RST parsing, rhetorical graph construction, and planning; *Generator* refers to the model used for final answer generation. 8B = Llama-3.1-8B; 70B = Llama-3.3-70B.

The mixed configuration (8B structurer + 70B generator) achieves an LLM Score of 60.52, recovering the majority of the gain obtained by the full 70B setting (62.07) and substantially outperforming standard RAG (49.33). Even the all-8B variant reaches 58.94, indicating that smaller models can produce discourse structures that still benefit downstream generation. The remaining gap between the 8B+70B and all-70B settings (1.55 LLM Score) suggests that higher-quality structural signals from larger models still contribute incrementally, yet the cost reduction from offloading structural inference to an 8B model may be preferable in resource-constrained deployments. These results confirm that the modular architecture of Disco-RAG supports flexible allocation of model capacity across components without forfeiting the core benefits of discourse-aware generation.

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 the 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 \uparrow	SummaC \uparrow
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 7.8: Impact of supervised fine-tuning (SFT) and discourse conditioning.

All systems share the same retrieval pipeline and decoding configuration, and we report RL and SummaC on the SciNews test set. Table 7.8 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.

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 $p < 0.05$.

Human Evaluation. We conduct a human evaluation on the SciNews dataset. We randomly sample 15 test articles and ask three graduate students 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. (2024b), 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 0.73, 0.65, 0.66, and 0.68 on the four dimensions, indicating substantial consistency among annotators. Table 7.9 reports the average scores across all annotated samples. Detailed instructions for the human raters are provided in Section E.4.

System	Relevance \uparrow	Simplicity \uparrow	Conciseness \uparrow	Faithfulness \uparrow
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 7.9: Average human ratings on SciNews. Scores are computed on a three-point Likert scale, and higher values indicate better performance.

Table 7.9 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.

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 7.5, Figure 7.6, and Figure 7.7.

Case Study 1

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

[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 7.5: 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?*

[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 7.6: 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?*

[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 7.7: 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.

7.5 Related Work

Structure-Aware Retrieval-Augmented Generation. Retrieval-Augmented Generation (RAG) enhances LLMs in knowledge-intensive tasks by retrieving external evidence (Lewis et al. 2020b). 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; Zhou et al. 2026) 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 et al. 2025; Li et al. 2025b), or employ alternative formats like hierarchical graphs (Wang et al. 2025c; Huang et al. 2025a), semantic chunking (Wang et al. 2025d; Qu et al. 2025; Zhao et al. 2025), trees (Sarthi et al. 2024), and tables (Lin et al. 2025). More adaptive strategies dynamically select structures based on context (Li et al. 2025c). 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.

Rhetorical Structure Theory for Text Generation. Rhetorical Structure Theory (RST; Mann et al. 1987; Mann et al. 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; Marcu. 1999a; Mann et al. 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. 2023c; Liu et al. 2024a). 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 structure.

7.6 Chapter Summary

This chapter addressed Research Question 5 by investigating whether discourse structure can be leveraged to organize structurally disconnected retrieved evidence in retrieval-augmented generation. The proposed Disco-RAG framework augments standard RAG pipelines with discourse-aware modeling at three levels, namely local structure within chunks, global rhetorical relations across chunks, and a planning stage that synthesizes these signals into an ordered generation blueprint. Experiments on Loong, ASQA, and SciNews demonstrated that this approach yields consistent improvements in answer quality and factual consistency over both standard RAG baselines and prior state-of-the-art methods, without requiring in-domain fine-tuning. Ablation studies confirmed that all three structural components contribute complementarily, with the intra-chunk RST tree and inter-chunk rhetorical graph proving most critical in long-context settings. Controlled perturbation experiments further showed that the observed gains depend on the quality of the discourse structures rather than on the mere presence of additional structural tokens, providing empirical support for Hypothesis V.

With this chapter, the dissertation completes a trajectory that began with distributional discourse representations for attention modulation in long-document encoding (Chapter 3), extended through parameter-efficient adaptation under constrained update budgets (Chapter 4), advanced to macro-level planning for multimodal summarization (Chapter 5), progressed to fine-grained control over explanatory content in public-facing summaries (Chapter 6), and arrived at the organization of retrieved evidence through discourse structure in retrieval-augmented generation (Chapter 7). Across these settings, the recurring finding is that making organizational decisions explicit, whether as soft attention biases, structured intermediate plans, or rhetorical scaffolds over retrieved passages, tends to yield improvements in controllability and factual reliability that are difficult to obtain through implicit end-to-end learning alone. The following chapter synthesizes these findings, discusses limitations, and outlines directions for future research.

8

Conclusion

This dissertation investigates how discourse structure can inform and improve text generation across a variety of challenging scenarios, spanning long-document summarization, multimodal and explanatory summarization, and retrieval-augmented generation. The central thesis is that explicit modeling of rhetorical organization, whether through distributional representations of discourse relations, structured generation plans, or rhetorical scaffolds over retrieved evidence, enables generation systems to make better decisions about content selection, ordering, and presentation. Five studies are presented to examine this thesis from complementary angles. [Chapter 3](#) and [Chapter 4](#) focus on integrating discourse information into the encoding process for long-document summarization. [Chapter 5](#) and [Chapter 6](#) extend this perspective to the generation stage through explicit planning mechanisms in multimodal and explanatory summarization settings. [Chapter 7](#) broadens the scope by applying discourse structure modeling to retrieval-augmented generation, where retrieved evidence lacks explicit rhetorical connections, and the generator operates over structurally disconnected passages.

This concluding chapter synthesizes the main findings and reflects on their broader implications. [Section 8.1](#) summarizes the primary contributions of each study. [Section 8.2](#) distills cross-cutting insights that emerge from the collective evidence and discusses what these findings reveal about the role of discourse in text generation. [Section 8.3](#) acknowledges the limitations of the current work, and [Section 8.4](#) outlines promising directions for future research. The use of AI tools is disclosed in [Section 8.5](#), and ethical considerations ([Section 8.6](#)) are discussed to address the responsible use of data and models throughout this dissertation.

8.1 Core Contributions

This dissertation applies discourse as a guiding perspective to study text generation from complex or distributed evidence. Beyond surface fluency, the central interest lies in whether explicitly conditioning on discourse structure can help generation systems make more robust and consistent decisions about *what to include*, *how to order evidence*, and *how to organize the output* when inputs become longer, adaptation budgets are limited, the modality and audience change, or the evidence is retrieved from external sources without pre-existing rhetorical connections. The main contributions of this dissertation are summarized below.

- **Discourse distributions for long-document summarization.** Chapter 3 proposes an uncertainty-aware formulation of discourse structure as distributions over rhetorical relation types, rather than relying on a single 1-best parse. These distributional signals are injected into long-context Transformer attention so that local information aggregation can be guided by relation semantics and salience asymmetries. This design supports a more fine-grained analysis of when relation labels and structural confidence influence summary coverage and discourse coherence.
- **Discourse-guided parameter-efficient adaptation.** Chapter 4 studies discourse signals under limited update budgets and introduces discourse-aware low-rank variants that make it possible to integrate discourse information without full model fine-tuning. It develops controlled comparisons between label-aware and label-agnostic encodings, and between uncertainty-preserving and deterministic encodings, to separate the role of discourse semantics from the role of structural confidence. The empirical analyses focus on behavioral consistency under constrained adaptation, while also considering how stronger structural control may interact with factual risk.
- **Discourse-aware planning for video summarization.** Chapter 5 extends the discourse perspective to scientific video-to-text summarization and introduces a large-scale dataset of aligned presentation videos and paper abstracts, together with a benchmarking protocol intended to support comparison. It then studies a discourse-aware planning generation pipeline in which an explicit organizational plan guides summary generation, motivated by the relatively regular discourse organization of scientific abstracts. The analyses examine how planning affects discourse-level coherence and factual grounding when multimodal evidence is incomplete or noisy.
- **Explanatory summarization for public audiences.** Chapter 6 treats explanation as

a discourse component that can be planned, located, and controlled, and it formulates planning as a sequence of questions that reflects how human writers introduce explanatory sentences. This representation supports more explicit control over where explanations appear and what they target, while keeping the overall summary structure interpretable. In addition, it also discusses evaluation mismatch in explanatory settings and motivates verification-oriented analyses alongside readability and coverage, in order to better separate helpful explanations from unreliable elaborations.

- **Discourse-aware retrieval-augmented generation.** Chapter 7 extends discourse structure modeling beyond summarization to retrieval-augmented generation, where retrieved passages arrive as structurally disconnected chunks. It proposes a framework that constructs local RST trees within each chunk, infers a directed rhetorical graph across chunks to capture inter-chunk relations, and introduces a discourse-driven planning module that synthesizes these structural signals into an ordered blueprint for answer generation. Ablation and perturbation analyses provide evidence that the observed improvements stem from the quality of the discourse structures rather than from the mere presence of additional structural tokens.

8.2 Synthesis of Findings

This dissertation presents five complementary studies that collectively investigate the role of discourse structure in text generation across a range of scenarios, from long-document summarization to retrieval-augmented question answering. The following discussions distill several core insights from these studies and offer broader reflections on discourse-aware generation methods.

Uncertainty as Informative Signal. Experiments in Chapter 3 and Chapter 4 consistently demonstrate that representing discourse structures as probability distributions, rather than committing to a single 1-best parse tree, leads to more robust summarization performance. While this finding may appear counterintuitive given the conventional expectation that parsers should output the most confident structural annotation, it aligns with a growing body of evidence that discourse relations are not always uniquely determined. Crowd-sourced annotation studies have shown that human readers often assign different but individually plausible relation labels to the same text span (Scholman et al. 2022), and that modeling these label distributions, rather than collapsing them into a

single category, can improve classification of implicit discourse relations (Yung et al. 2022). More recently, the Enhanced RST framework has formalized the notion that multiple rhetorical relations may hold concurrently at the same position in a discourse graph (Zeldes et al. 2025). These observations suggest that a forced single parse may discard informational value that is inherent in the structural ambiguity of the text. When summarization models have access to n-best discourse forests and their associated confidence scores, they can learn to dynamically weigh alternative structural interpretations depending on context, rather than being constrained by a potentially erroneous deterministic decision.

Explicit Structure and Implicit Learning. Although large-scale language models are capable of implicitly acquiring shallow discourse patterns from vast amounts of text, multiple experiments in this dissertation demonstrate that explicitly incorporating discourse structures yields measurable performance gains. One plausible explanation is that most pretraining objectives do not directly optimize for discourse-level organization. Masked language modeling and next-token prediction primarily capture local lexical co-occurrence and syntactic patterns rather than global structural relations that span multiple paragraphs. A notable exception is the Next Sentence Prediction (NSP) objective used in BERT-family models, which does encode a form of inter-sentential coherence. Shi et al. (2019) showed that BERT models trained with NSP achieve strong performance on implicit discourse relation classification, and BERT-based encoders remain competitive for this task to date (Gessler et al. 2021). However, NSP operates at the level of sentence pairs and does not capture the kind of document-level structural dependencies, such as long-range salience asymmetries and multi-paragraph rhetorical organization, that are central to the generation settings studied in this dissertation. In Chapter 3, annotated discourse distribution tensors are injected into Longformer’s sparse attention computations, while in Chapter 4, RST distribution matrices serve as modulation coefficients for the input representations of LoRA. Both approaches introduce structural signals through soft modulation and enable discourse knowledge to complement the model’s native linguistic capabilities without interference. Ablation studies in Chapter 4 further show that replacing real discourse distributions with random values causes performance to degrade to near-baseline levels. This confirms that the benefit of structural signals stems from their semantic content rather than from regularization effects or random perturbations.

Planning as Structured Generation Constraint. Chapter 5, Chapter 6, and Chapter 7 introduce planning as an intermediate stage in the generation process. Prior work has established that decomposing generation into a content selection step followed by surface realization can improve both faithfulness and controllability. Puduppully et al. (2019) demonstrated this principle for data-to-text generation, where entity-based content plans determine what information to convey and in what order. Narayan et al. (2021) extended this idea to abstractive summarization by prepending entity chains as learned prompts, showing that such plans reduce hallucinations by grounding the decoder in an explicit content outline. More recently, Narayan et al. (2023) proposed question-answer blueprints as a richer form of plan that captures not only what to say but also the communicative intent behind each summary sentence, and Adams et al. (2023) showed that EDU-level content plans can guide the generation of more diverse and higher-quality summary candidates for re-ranking. The findings in Chapter 5, Chapter 6, and Chapter 7 are consistent with this line of work and extend it in three directions. First, in the multimodal setting of Chapter 5, planning helps bridge the gap between temporally fragmented video evidence and the relatively regular rhetorical organization expected of scientific abstracts. Second, in the lay summarization setting of Chapter 6, planning at the EDU level enables fine-grained control over where and what type of explanatory content is generated, a capability that phrase-level plans do not readily support. Third, in the retrieval-augmented setting of Chapter 7, planning operates over discourse structures inferred from independently retrieved passages, synthesizing intra-chunk and inter-chunk rhetorical signals into an ordered blueprint that guides the generator toward more coherent and factually grounded answers. Controllable generation experiments further demonstrate that modifying the planning signals can reliably influence the structure and content of the output, whereas end-to-end models exhibit greater instability when presented with similar high-level instructions.

Relation Semantics for Content Selection. Chapter 3 and Chapter 4 compare model variants that either retain or omit discourse relation labels, with consistent results indicating that distinguishing among rhetorical relation types leads to greater performance gains. This finding aligns with a long-standing assumption in discourse-based generation. Early work by Marcu (1997) and Marcu (1999b) demonstrated that the nucleus-satellite distinction in RST trees can serve as an effective heuristic for content selection, while Louis et al. (2010) showed that discourse indicators are informative predictors

of sentence importance. More recent neural approaches have confirmed that integrating RST structure into extractive and abstractive models improves content selection (Xu et al. 2020; Xiao et al. 2020; Dong et al. 2021), though these studies typically rely on the 1-best tree without distinguishing among relation types. The results in this dissertation suggest that relation-type information provides semantic cues that go beyond structural centrality alone. For example, satellite units marked as *Elaboration* often contain detailed expansions of core propositions and may be more readily omitted during content compression, whereas units marked as *Evidence* or *Reason* serve argumentative functions and should be retained or excluded depending on whether the associated nucleus appears in the summary. Further experiments in Chapter 6 demonstrate that, in the context of lay-oriented science summarization, relations such as *Explanation* and *Background* are especially critical for producing informative and readable summaries. Removing planning questions associated with these relation types leads to notable declines in both informativeness and readability, reinforcing the view that relation semantics carry task-relevant information that coarser structural signals do not fully capture.

From Encoding to Planning Paradigms. The trajectory of this dissertation reveals a methodological evolution in how discourse information is utilized, moving from the encoding stage to the generation stage, and from implicit modulation to explicit planning. Earlier discourse-aware summarization systems typically inject structural information into the encoder, for example, by converting RST trees into dependency structures that modulate attention (Xiao et al. 2020), by incorporating discourse graphs into transformer encoders (Xu et al. 2020; Dong et al. 2021), or by using section-level discourse cues to reshape long-document attention patterns (Cohan et al. 2018). Chapter 3 and Chapter 4 follow this encoder-side strategy but extend it by preserving relation-type distributions and parsing uncertainty. This approach has the advantage of leaving the decoder architecture unchanged, but its influence on generation decisions remains implicit and difficult to interpret directly. In contrast, Chapter 5 and Chapter 6 adopt a planning paradigm that externalizes discourse organizational decisions into an explicit sequence of actionable questions. This makes structural information more transparent and traceable throughout the generation process. Chapter 7 extends this trajectory to the RAG setting by applying both RST-based structural analysis and discourse-driven planning to retrieved evidence. This extension demonstrates that the planning paradigm remains effective even when the discourse structures are not inherited from a source document

but are instead constructed over retrieved fragments. The shift across these chapters reflects a deeper understanding of discourse. Rather than serving merely as auxiliary cues for source comprehension, discourse structures are reframed as active frameworks for organizing the generated output, whether it takes the form of a summary or a retrieval-augmented response. The planning-based approach enables this organizational role to be explicitly modeled while also providing clearer interfaces for user intervention and system debugging.

Evaluation Gaps and Human–Machine Disparity. Although the methods proposed in this dissertation achieve consistent improvements on a range of automatic evaluation metrics, human evaluation results indicate that a substantial gap remains between machine-generated texts and those written by humans. This gap is particularly pronounced with respect to factual accuracy, the usefulness and accuracy of explanations, and discourse coherence. Notably, existing automatic metrics often fail to capture these deficiencies, an observation that echoes broader concerns in the generation evaluation literature. [Cao et al. \(2022a\)](#) demonstrated that a substantial portion of so-called hallucinations in machine-generated texts are in fact consistent with world knowledge and can be beneficial for readers, yet standard consistency detectors penalize them uniformly. Analyses in [Chapter 6](#) confirm this pattern in the lay summarization setting, where conventional factual consistency detectors incorrectly flag helpful human-written explanations as hallucinations, precisely because such explanations introduce background knowledge that is not explicitly present in the source document. The modified evaluation protocol introduced in [Chapter 6](#), which incorporates external knowledge verification via retrieval, represents one step toward addressing this mismatch, but a more fundamental challenge remains. When generation tasks require the incorporation of content that extends beyond the source, consistency metrics grounded solely in source-document entailment become inadequate. Defining and operationalizing the notion of answer usefulness requires accounting for the target audience’s prior knowledge and informational needs, which are factors that remain largely unaddressed by current automatic evaluation methods.

8.3 Research Limitations

Data. All datasets employed in this dissertation are publicly available and subject to peer review. Nonetheless, they reflect only a limited subset of real-world generation scenarios. The corpora span diverse domains, including books, scientific publications, legal documents, conference presentations, and knowledge-intensive QA benchmarks. However, potential biases in content and label distributions have not been systematically examined. Furthermore, the datasets are predominantly in English, and the proposed methods have not been evaluated on multilingual benchmarks or other languages.

Parser. The discourse-aware methods developed in this dissertation depend on RST parsers to produce structural annotations. In [Chapter 3](#) through [Chapter 6](#), trained parsers are employed, while [Chapter 7](#) adopts a zero-shot LLM-based parser. In both cases, parser accuracy may deteriorate for longer or out-of-domain inputs, potentially leading to error propagation in downstream tasks. Given the high cost of obtaining manually annotated RST trees at scale, direct comparisons between parser outputs and gold-standard annotations remain limited. As such, the extent to which improvements in parsing accuracy could enhance downstream generation quality remains an open question.

Model. This dissertation utilizes a variety of open-source language models, including Longformer, Vicuna, Mistral, Llama-3, Qwen2.5, and several video-based large multi-modal models. While consistent improvements are observed across these backbones, the number of possible architectures and parameter scales is vast, and not all configurations have been explored. Moreover, closed-source models such as GPT-4 and GPT-o1 are employed for evaluation and plan generation. However, the substantial costs associated with API usage constrain the breadth of experiments involving these models.

Automated Evaluation. This dissertation applies a suite of widely used metrics, including ROUGE, BERTScore, SummaC, and several task-specific measures, to assess generation quality from multiple perspectives. Nonetheless, these metrics have known limitations and may not fully capture all dimensions of output quality. Additionally, evaluations using large language models (LLMs) are incorporated, but such evaluators may exhibit biases inherited from pretraining and often lack interpretability.

Human Evaluation. Due to the considerable length of source input, it is non-trivial to scale human evaluation via crowd-sourcing. As a result, each study includes a limited number of annotated samples that may not be representative of the full dataset. While all annotators are graduate students with backgrounds in computer science or computational linguistics, they may not possess domain expertise across all evaluation contexts. Thus, their assessments should be interpreted in conjunction with automated evaluation results.

Efficiency. Incorporating discourse structure into the generation pipeline introduces additional computational overhead. In [Chapter 3](#) and [Chapter 4](#), RST parsing is applied as an offline preprocessing step, and the runtime cost during training and inference remains comparable to the baseline. In [Chapter 7](#), however, the framework requires additional LLM calls for intra-chunk RST parsing, inter-chunk rhetorical graph construction, and discourse-driven planning, which increases both token consumption and end-to-end latency. While the overhead is moderate under the experimental settings reported in this dissertation, deploying discourse-aware methods in latency-sensitive or large-scale applications would benefit from engineering optimizations such as caching pre-parsed discourse structures, batching structural queries, or distilling lighter parsing and planning modules. The trade-off between structural richness and runtime efficiency remains an open consideration across the methods proposed in this dissertation.

Scope. The methods proposed in this dissertation focus on summarization and retrieval-augmented generation. Their applicability to other natural language processing tasks, such as machine translation or dialogue generation, has not been explored. While the discourse-aware mechanisms could potentially be adapted to these tasks, such extensions remain directions for future research.

8.4 Future Directions

Advancing Discourse Parsing and Joint Learning. The discourse-aware methods in this dissertation rely on external RST parsers to generate structural annotations. This makes downstream performance sensitive to parser accuracy. A direction for future research is to develop joint learning frameworks that integrate discourse parsing and downstream generation and allow structural annotations to be optimized end-to-end for the target task rather than relying on fixed, pretrained parsers. Moreover, while this dis-

sertation primarily adopts Rhetorical Structure Theory and Question under Discussion as the discourse representation framework, alternative theories such as the Penn Discourse Treebank (PDTB) relation inventory could be explored to assess their impact on generation quality. Such comparisons may help clarify the relative advantages of different discourse formalisms in capturing long-range dependencies and organizing content.

Extending to Multilingual Settings. All experiments in this dissertation are conducted on English datasets. Extending discourse-aware generation methods to other languages is both technically and theoretically feasible, but it introduces several challenges. Discourse structures may vary significantly across languages in terms of rhetorical relation distributions, nucleus–satellite annotation conventions, and intersentential connectives. Moreover, many languages lack large-scale annotated discourse corpora, which poses challenges for parser transferability. Future work may explore cross-lingual transfer learning that uses multilingual pretrained models and limited annotation. This would enable the application of the proposed methods in low-resource language settings.

Fusion with Broader Modalities. This dissertation extends discourse-aware planning from purely textual to the video domain and demonstrates its effectiveness in multimodal contexts. Future research could explore integration with additional modalities, such as scientific documents that contain charts, formulas, and tables; combinations of presentation slides and spoken narration; or hybrid inputs that involve video, audio, and transcripts. These scenarios require novel cross-modal discourse representations capable of modeling the semantic roles of visual and textual elements. For example, figures in scientific papers often function as evidential or explanatory components. Incorporating their discourse roles into planning frameworks poses an interesting challenge. Furthermore, when different modalities offer conflicting or redundant information, effective strategies for planning-level coordination and content selection become essential.

Task Generalization Beyond Summarization and RAG. This dissertation applies discourse-aware methods to summarization and retrieval-augmented generation. The underlying mechanisms may benefit a broader range of NLP tasks that require long-range reasoning or structured content generation. In document-level machine translation, discourse information can guide cross-sentence consistency. In dialogue generation, discourse planning may facilitate the maintenance of thematic coherence across

turns. Adapting the proposed planning framework to such tasks would require redefining the plan format and aligning it with task-specific optimization objectives.

Interactive and Controllable Generation. Planning-based methods offer inherent controllability because modifying the plan directly influences the structure and content of the generated output. Our experiments demonstrate that adjusting the planning questions allows control over output length, readability, and explanation types. Building on this, future work may develop interactive generation systems that enable users to refine or customize outputs by editing, adding, or removing planning questions. Such human-in-the-loop workflows hold promise for applications in science communication, educational content creation, and personalized information access. Moreover, incorporating implicit user feedback into planning updates could allow the system to adaptively improve its outputs based on user preferences.

Deepening the Integration of Discourse and Retrieval. Chapter 7 takes a first step toward integrating discourse structure modeling with retrieval-augmented generation by constructing RST trees and rhetorical graphs over retrieved passages. However, the current framework applies discourse analysis after retrieval, treating the retriever and the discourse parser as independent modules. A promising direction for future work is to tighten this coupling so that discourse signals inform the retrieval process itself. For example, a discourse-aware retriever could prioritize passages that fill specific rhetorical roles relative to already retrieved evidence, such as selecting a passage that provides a contrasting view or an elaboration on a partially supported claim. Additionally, the current framework relies on a fixed retrieval step followed by a single round of discourse analysis and planning. Iterative approaches that alternate between retrieval, discourse structure construction, and plan refinement could allow the system to adaptively seek additional evidence when the rhetorical scaffold reveals gaps in the argument. Exploring how discourse-level feedback signals can guide retrieval updates represents an open challenge at the intersection of structured reasoning and information retrieval.

Improving Evaluation Methodologies. Widely used automatic evaluation metrics often fail to capture discourse coherence and output usability. Our attempt to use external knowledge validation as a proxy for factual consistency represents a preliminary step, but limitations remain. Future work may develop more fine-grained evaluation proto-

cols that can assess logical coherence among arguments and measure comprehensibility for the target audience. Incorporating discourse structure into factual consistency evaluation is also a promising direction. For example, RST relations could be leveraged to assess whether generated content is supported by the source evidence. Developing automatic evaluation metrics that better align with human judgment would enhance the reliability of benchmarking across generation tasks.

8.5 The Use of AI Tools

During the preparation of this dissertation, AI-based writing assistants (e.g., GPT-5 and Claude) are employed solely for language polishing, grammar correction, and stylistic refinement. These tools play no role in research ideation, methodology design, experimental execution, data analysis, or the interpretation of results. All scientific content, analytical reasoning, and conclusions presented herein are conceived, developed, and validated entirely by the author, who assumes full responsibility for the accuracy and integrity of this work.

8.6 Ethical Considerations

All datasets employed throughout this dissertation are publicly accessible and adhere to their original licenses and usage policies. For datasets collected as part of this dissertation, including VISTA and SciNews, appropriate permissions have been obtained, and the collection process strictly follows platform terms and relevant copyright regulations for academic research. Throughout the stages of data processing, experimental analysis, and model training and evaluation, our pipeline operates entirely on de-identified text without collecting or inferring personal identities or sensitive attributes, and no instances of privacy infringement have been identified.

All human evaluations in this dissertation follow voluntary participation protocols with appropriate compensation. Participants are recruited from master's or doctoral students with backgrounds in computer science or computational linguistics, and all possess advanced proficiency in English. This dissertation complies with [ACL Policy on Publication Ethics](#), and we caution against applying our systems in high-stakes environments without additional safeguards and human oversight.

Appendices



Appendix for Chapter III

A.1 Questionnaire of Human Evaluation

- **Faithfulness**

1. Completely hallucinated content
2. A lot of hallucinated content and factual mistakes
3. Most content is supported by the source document
4. Only one or two characters or events contradicted or not mentioned in the source
5. All information in the summary is faithful/supported by the source

- **Informativeness**

1. No important information in the source is covered in the summary
2. Only covers a small fraction of the source document information; one cannot learn the main content of the story from only the summary
3. Covers around half of the important points from the source; one can learn the main content of the story from only the summary
4. Only a few important points are missing in the summary
5. All important information is summarized

- **Readability**

1. Not understandable at all
2. Hard to understand the content of the summary
3. The summary is overall readable, with most sentences correct and fluent

4. Easy to understand, with only occasional grammatical mistakes or incoherent sentences
5. Fluent, with minor or no grammatical mistakes, coherent sentences, and clear structure

- **Conciseness**

1. All information in the summary is redundant or unimportant
2. Most of the information in the summary is redundant or unimportant
3. Around half of the content in the summary is redundant
4. Only a few points in the summary are redundant
5. No information in the summary is redundant

B

Appendix for Chapter IV

B.1 GPT-4 Prompts

Zero-shot Summaries Generation

Document: {Document}
Summary:

Figure B.1: Prompt for zero-shot summaries generation.

In-context Summaries Generation

Document: {Document}
Summary: {Summary}

Document: {Document}
Summary: {Summary}

Document: {Document}
Summary:

Figure B.2: Prompt for in-context summaries generation.

B.2 Human Evaluation Guidelines

Summaries Evaluation

Source Document: {Document}
 Summary of Candidate1: {Candidate1}
 Summary of Candidate2: {Candidate2}
 Summary of Candidate3: {Candidate3}
 Summary of Candidate4: {Candidate4}
 Summary of Candidate5: {Candidate5}

Note: The summaries are presented in order, with their respective candidate numbers from 1 to 5.

Please review the following evaluation guidelines to assess the quality of the above five candidate summaries, and rank them from best to worst:

Evaluation Guidelines: {Guidelines}

Please use the following format for your output (scores ONLY):

Relevance of Candidate1:
 Informativeness of Candidate1:
 Conciseness of Candidate1:
 Faithfulness of Candidate1:
 Relevance of Candidate2:
 Informativeness of Candidate2:
 Conciseness of Candidate2:
 Faithfulness of Candidate2:
 Relevance of Candidate3:
 Informativeness of Candidate3:
 Conciseness of Candidate3:
 Faithfulness of Candidate3:
 Relevance of Candidate4:
 Informativeness of Candidate4:
 Conciseness of Candidate4:
 Faithfulness of Candidate4:
 Relevance of Candidate5:
 Informativeness of Candidate5:
 Conciseness of Candidate5:
 Faithfulness of Candidate5:
 Ranking (using candidate number):

Figure B.3: Prompt for summaries evaluation.

Prerequisites To participate in this evaluation, you must meet the following two criteria: (1) be a Master’s or Ph.D. student in Computer Science or Computational Linguistics, and (2) demonstrate English proficiency at C2 level or higher.^a If you do not meet both criteria, we kindly ask you to refrain from participating in this task. Eligible participants are encouraged to follow the instructions below carefully.

Instructions We invite you to carefully review the following long document along with five candidate summaries. After a thorough examination of each summary, please rate them based on the following four criteria, using a Likert scale from 1 (worst) to 5 (best), where a higher score denotes better quality:

- **Relevance:** This metric assesses the extent to which the summary content accurately reflects the source text. A relevant summary should encompass topics pertinent to the source document.
- **Informativeness:** This metric assesses the extent to which the summary provides a comprehensive understanding of the key points and essential details from the source text. An informative summary should encapsulate the core ideas, facilitating a clear and precise comprehension of the main arguments and findings of the source document.
- **Conciseness:** This metric assesses the extent to which the summary excludes less important information from the source text. A concise summary should effectively eliminate non-essential content from the source document during the generation process.
- **Faithfulness:** This metric assesses the extent to which the candidate is incorrect in that it contradicts the information from the source document. A faithful summary adheres strictly to the information provided in the source document, avoiding the inclusion of unverified facts.

Overall Ranking After assigning scores to each summary for the individual criteria, rank all candidates from best to worst based on their overall quality.

^a https://en.wikipedia.org/wiki/C2_Proficiency

Figure B.4: A snapshot of the experimental instructions provided to human evaluators.

C

Appendix for Chapter V


C.1 Data Sample

The VISTA dataset contains carefully curated video-text pairs, predominantly sourced from published papers, aiming to ensure a high standard of quality and relevance. The accompanying texts are designed to function as summaries of their respective videos, offering a concise representation of their content (see [Figure C.1](#)). Additionally, our dataset focuses on topics within the field of artificial intelligence, making it a good resource for research in AI-related video-to-text summarization and comprehension.

C.2 Prompts Used in Our Study





C.3 Human Evaluation Guidelines

C.4 Prompt for GPT-o1 to Evaluate Summary Quality



When Does Translation Require Context? A Data-driven, Multilingual Exploration

Patrick Fernandes*, Kayo Yin*, Emmy Liu
André F. T. Martins, Graham Neubig

* equal contribution

Although proper handling of discourse significantly contributes to the quality of machine translation (MT), these improvements are not adequately measured in common translation quality metrics. Recent works in context-aware MT attempt to target a small set of discourse phenomena during evaluation, however not in a fully systematic way. In this paper, we develop the Multilingual Discourse-Aware (MuDA) benchmark, a series of taggers that identify and evaluate model performance on discourse phenomena in any given dataset. The choice of phenomena is inspired by a novel methodology to systematically identify translations that require context. This methodology confirms the difficulty of previously studied phenomena while uncovering others which were not previously addressed. We find that commonly studied context-aware MT models make only marginal improvements over context-agnostic models, which suggests these models do not handle these ambiguities effectively. We release code and data for 14 language pairs to encourage the MT community to focus on accurately capturing discourse phenomena.

Figure C.1: A random sample from the VISTA dataset, originating from [Fernandes et al. \(2023\)](#).

Quality Control Guidelines

Evaluate each video-text pair to determine whether the text provides a concise and accurate summary of the corresponding video.

- **Concise:** Ensure the text is brief, focused, and free of unnecessary details.
- **Accurate:** Verify that the text faithfully represents the video's content.

Make binary judgments (Valid or Invalid) for each pair. If flagged as Invalid, provide a brief justification.

Answer:

Judgment: (Valid or Invalid)

Justification: (Justification if flagged as invalid)

Figure C.2: Quality control guidelines.

Summary Generation (without plan)

Generate a summary for the provided content.

Content: {Video/Audio/Transcript/OCR}

Summary:

Figure C.3: Prompt to generate summaries without plans.

Question Generation

Generate a coherent and contextually relevant question based on the provided context and target sentence, ensuring that the target sentence can be treated as an answer to the generated question.

Context: {Context Text}

Target: {Target Sentence}

Question Sentence:

Figure C.4: Prompt for question generation.

Prompt for PG model

Generate a list of questions for the provided {Video/Audio/Transcript...}.

Content: {Video/Audio/Transcript...}

Questions:

Figure C.5: Prompt for PG model.

Prompt for SG model

Generate a summary for the following {Video/Audio/Transcript...} based on the plan questions.
 Content: {Video/Audio/Transcript...}.
 Plan Questions: {Questions}
 Ensure that the generated summary sequentially answers the plan questions.
 Summary:

Figure C.6: Prompt for SG model.

Irrelevant Question Generation

Randomly generate a question with a question mark.
 Question Sentence:

Figure C.7: Prompt used by GPT-o1 to generate irrelevant questions.

Summary Readability Modification

Rewrite the following text to further adjust the style or detail.
 Here is the text to be rewritten: {Text}
 Refine the above text to be more {lay/expert} style.
 Modified Text:

Figure C.8: Summary readability modification.

Summary Length Modification

Rewrite the following text to further adjust the style or detail.
 Here is the text to be rewritten: {Text}
 Shorten the above text by about {10% / 30% / 60%}. Focus on the key points and remove less critical details.
 Modified Text:

Figure C.9: Summary length modification.

Plan Readability Modification

Rewrite the following questions to further adjust the style or detail.
 Here are the questions to be rewritten:
 1. {Q1}
 2. {Q2}
 ...
 Refine the above questions to be more {lay/expert} style.
 Modified Questions:

Figure C.10: Plan readability modification.

Prerequisites To participate in this evaluation, you must meet the following two criteria: (1) be a Master’s or Ph.D. student in Computer Science or Computational Linguistics, and (2) demonstrate English proficiency at C2 level or higher.^a If you do not meet both criteria, we kindly ask you to refrain from participating in this task. Eligible participants are encouraged to follow the instructions below carefully.

Instructions The following section provides detailed descriptions of the evaluation metrics and criteria used in this study. Please review the accompanying source video and the candidate summaries thoroughly. After evaluating each summary, assign scores based on the five criteria below, using a 1-to-5 Likert scale where higher scores indicate better quality:

- **Faithfulness:** Assess the accuracy of the summary in representing the content of the source video. A faithful summary should adhere closely to the source material, avoiding contradictions, misinterpretations, or unverified information.
- **Relevance:** Measure how well the summary includes the topics and themes central to the source video. A relevant summary should focus on the content that is most pertinent to the original video.
- **Informativeness:** Evaluate the extent to which the summary captures the main points and essential details of the source video. An informative summary should provide a clear and comprehensive understanding of the video’s core ideas and findings.
- **Conciseness:** Determine the efficiency of the summary in conveying information. A concise summary should avoid redundancy and extraneous details while retaining all critical information from the source video.
- **Coherence:** Examine the logical flow and overall structure of the summary. A coherent summary should present information in an organized and easy-to-follow manner, ensuring that ideas connect naturally and transitions between points are smooth.

Rating System For each metric, use the following Likert scale:

- 1 (Worst): Does not meet the criteria at all.
- 2 (Poor): Meets the criteria minimally.
- 3 (Fair): Meets the criteria adequately.
- 4 (Good): Meets the criteria well.
- 5 (Best): Fully meets the criteria.

Overall Ranking After assigning scores to each summary for the individual criteria, rank all candidates from best to worst based on their overall quality. Consider the summaries’ performance across all criteria when determining the final rankings.

^a https://en.wikipedia.org/wiki/C2_Proficiency

Figure C.11: A snapshot of the experimental instructions provided to human evaluators.

Source Video: {Source Video}

Candidate Summary: {Candidate Summary}

You are tasked with evaluating the quality of the candidate summary based on the provided source video. Please adhere strictly to the following evaluation guidelines and scoring criteria to ensure a consistent and objective evaluation.

Evaluation Guidelines: {Guidelines}

Instructions for Output:

- Provide your evaluation using the following format, outputting scores only.
- Assign a score from 1 to 5 for each dimension, with 1 being the lowest and 5 being the highest.

Output Format:

- Faithfulness: [Score]
- Relevance: [Score]
- Informativeness: [Score]
- Conciseness: [Score]
- Coherence: [Score]

If you encounter ambiguity in evaluating any dimension, prioritize adherence to the evaluation guidelines and provide the most accurate score possible based on the provided information. Do not include any additional comments or justifications in your response.

Figure C.12: Prompt for GPT-o1 to evaluate summary quality.

D

Appendix for Chapter VI

D.1 Rule-based Extraction of Explanations

We develop a set of rules aimed at identifying words and phrases that indicate four types of explanations, which are in turn based on RST relations:

- **Background** includes *Background* and *Circumstance* relations, signalled by words such as “historically” and “traditionally.”
- **Elaboration** covers various *Elaboration* relations (e.g., *Additional*, *General-Specific*, *Part-Whole*, *Process-Step*, *Object-Attribute*, and *Set-Member*), as well as *Example* and *Definition* relations, signalled by words like “defined as” and “for example.”
- **Explanation** includes *Evidence*, *Explanation-Argumentative*, and *Reason* relations, signalled by words such as “because” and “due to.”
- **Comparison** includes *Comparison*, *Preference*, *Analogy*, and *Proportion* relations, signalled by words like “compared to” and “similarly.”

Table D.1 presents the words and phrases associated with each of these relations. To identify these indicators within the text, we compile regular expression patterns tailored to each category. Explanatory content is collected on a sentence-by-sentence basis: If a signal word is detected within a sentence, that entire sentence is categorized accordingly. In constructing these patterns, we use the caret symbol \wedge before certain words to restrict their occurrence to sentence-initial positions. This method ensures that words like “historically” or “similarly” are recognized only when they function as introductory elements, which is common in explanations. Additionally, we utilize regular expressions

such as `differ(?|ed|ing)` to capture different word forms. Once each explanatory sentence is identified, we then select the preceding sentence as its target sentence.

Category	Signal Words and Phrases
Background	^historically, ^traditionally, ^previously, ^in the past, ^before, ^initially, ^once, ^earlier, ^in the beginning, ^at first, ^prior to, ^originally, ^at the outset, ^at the time, ^long ago, ^decades ago, ^in former times, ^previously mentioned, ^the history of, ^the origin of, ^in earlier times, ^from the outset, ^in the early days, ^over the years, ^long before, ^centuries ago, ^during the early stages, ^at that time, ^back then, ^once upon a time, ^throughout history, ^previously established, ^over the course of history, ^in ancient times
Comparison	compared to, compared with, ^similarly, likewise, in contrast, in comparison, in opposition, ^on the contrary, ^on one hand, ^on the other hand, ^conversely, rather than, different from, ^unlike, similar to, analogous to, contrary to, in contradistinction, distinct from, distinguishable from, as opposed to, in the same way, by comparison, comparable to, differ(? :s ed ing) from, diverg(e es ed ing)? from, in a similar manner, in the same vein, on the flip side, correspondingly, on a different note, in opposition to, different from, in a contrasting way, in an analogous way
Elaboration	defined as, refer(? :s red ring)? to, mean(? :s t)?, known as, definition, ^in other words, ^that is to say, ^that's to say, ^this is to say, ^that means, ^this means, ^this implies, ^that implies, i.e., e.g., for example, for instance, such as, ^to clarify, ^to explain, whereas, ^to illustrate, ^to elaborate, ^specifically, ^particularly, in particular, as an example, by way of example, more precisely, ^to be specific, ^to exemplify, namely, by way of illustration, expounded upon, in more detail, one example, an example, ^to add to this
Explanation	because, due to, since, thanks to, owing to, for the sake of, stemming from, given that, in light of, for this reason, for that reason, for the reason that, for the purpose of, for this cause, the reason is, the reasons are, as a result, consequently, as a consequence, accordingly, with the result that, so that, such that, result(? :s ed ing)? in, , result(? :s ed ing)? from, lead(? :s ed ing)? to, which means, thereby, whereby, in consequence of, on account of, so as to, on the grounds that

Table D.1: Words and phrases used to heuristically identify discourse relations signaling explanations.

Question Generation Prompt

Generate a question that is coherent and contextually relevant to the provided context and target sentence.
 Context: {Context Text}
 Target: {Target Sentence}
 Question Sentence:

Figure D.1: Prompt used by GPT-4o and Mistral to generate plan questions based on target sentences and their context.

D.2 Human Expert Evaluation Results

Our human evaluators identify six examples where they could not make a decision on the accuracy and usefulness of explanations. Three are generated by GPT-4o_{ZS}, two by Mistral_{FT}, and one by Blueprint_{MT}, covering fields such as Biology, Neuroscience, and Materials Science. To further assess these cases, we enlist experts with a Ph.D. degree in the relevant areas. Using the same guidelines, GPT-4o receives an average rating of 2.33 for explanation accuracy and 2.67 for explanation usefulness. For Mistral_{FT}, the average rating is 3.50 for both accuracy and usefulness, while Blueprint_{MT} receives a rating of 4.00 for accuracy and 2.00 for usefulness. These expert evaluations align with the initial human assessment, reinforcing the consistency of the findings across the highlighted explanations.

D.3 Model Prompts

Summary Generation Prompt

Generate a lay summary for the provided document.
 Document: {Document Text}
 Lay Summary:

Figure D.2: Prompt used by Mistral_{ZS}, GPT-4o_{ZS} and Mistral_{FT} to generate summaries.

D.4 Human Evaluation Guidelines

Summary Generation Prompt (ICL)

Document: {Document Text}
 Lay Summary: {Summary Text}
 Using the example above as a reference for structure and tone, generate a new lay summary for the following document. Ensure the summary is original and does not replicate phrases or content from the example.
 Document: {Document Text}
 Lay Summary:

Figure D.3: Prompt used by *Mistral_{ICL}* and *GPT-4o_{ICL}* to generate summaries.

Named Entity Identification

Identify all named entities in the following text and return them as a list.
 Document: {Document Text}
 Return Format: [Entity 1, 2, 3]
 List all identified entities in the order they appear in the text.
 Identified Entities:

Figure D.4: Prompt used by *GPT-4o* to identify named entities for Blueprint model.

Plan-based Summary Generation Prompt

Generate a lay summary for the following document based on the plan questions.
 Document: {Document Text}
 Planning Questions: {Questions}
 Ensure that the generated summary sequentially answers the plan questions.
 Lay Summary:

Figure D.5: Prompt used by *Plan-Input* and *Plan-Output* models.

Direct Extraction of Explanatory Content

Parse the provided document to extract all explanatory sentences and the corresponding target sentences they explain. Return the parsing results as a list of dictionaries formatted as follows:

```
[
  {
    explanatory_sentence: 'XXX',
    target_sentence: 'XXX'
  },
  {
    explanatory_sentence: 'XXX',
    target_sentence: 'XXX'
  }
]
```

Ensure that each dictionary accurately pairs each explanatory statement with its respective target sentence.
Document: {Document Text}

Figure D.6: Prompt used by GPT-4o and Mistral to directly extract explanatory content.

Irrelevant Question Generation

Randomly generate a question with a question mark.
Question Sentence:

Figure D.7: Prompt used by GPT-4o to generate irrelevant questions.

Prerequisites Eligibility for this evaluation requires simultaneous fulfillment of two conditions: (1) being a Master’s or Ph.D. student in Computer Science or Computational Linguistics, and (2) demonstrating greater than or equal to C2 English proficiency^a. If you do not meet both criteria, we respectfully ask you to refrain from participating in this task. Those who qualify are encouraged to proceed and follow the instructions below.

Instructions Below is a detailed explanation of the metrics and evaluation criteria for our human evaluation process. Please carefully read the provided document along with the candidate summaries. After thoroughly examining each summary, evaluate them based on the following six criteria using a Likert scale from 1 to 5, with higher scores reflecting better quality:

- **Faithfulness:** This metric evaluates how accurately the summary reflects the information in the source document. A faithful summary should strictly adhere to the source material, avoiding any contradictions or unverified details.
- **Relevance:** This metric assesses how accurately the summary content reflects the topics covered in the source text. A relevant summary should include topics that are pertinent to the source document.
- **Informativeness:** This metric measures how well the summary conveys the key points and essential details from the source text. An informative summary should capture the main ideas, providing a clear and precise understanding of the source document’s arguments and findings.
- **Accessibility:** This metric evaluates how easy the summary is to read and understand. An accessible summary should be well-structured and written in clear language. It should avoid unnecessary complexity and ensure that readers can follow the content without difficulty.
- **Explanation Accuracy:** This metric measures how factually correct the explanatory content is. Explanation accuracy requires that the information provided in the explanatory sentence is verifiable and aligns with established knowledge in the relevant field, avoiding any misleading or incorrect explanations.
- **Explanation Usefulness:** This metric assesses how helpful the explanatory content is in enhancing the reader’s understanding of the subject matter. Explanation usefulness ensures that the content contributes to the reader’s comprehension of complex ideas, offering valuable insights and clarifications.

Please note that you are required to highlight all explanatory content found in the summary before assigning evaluation scores. You are permitted to use external knowledge sources, such as books, academic papers, and Wikipedia, but you are not allowed to use any AI tools to assist in your judgment. If you encounter any specific explanatory content that you are unable to evaluate for accuracy or usefulness, you should skip that content and mark it (meaning it will not factor into your scoring, but you will still need to score based on the rest of the content). This will allow us to bring in external experts (PhD holders in the relevant field) to assess it.

Rating System For each criterion, use the Likert scale as follows:

- 1 (Worst): Very poor quality, does not meet the criteria at all.
- 2 (Poor): Subpar quality, meets the criteria to a minimal extent.
- 3 (Fair): Average quality, adequately meets the criteria.
- 4 (Good): Above average quality, meets the criteria well.
- 5 (Best): Excellent quality, fully meets the criteria.

Overall Ranking After rating the summaries based on the six criteria, you are also expected to rank the candidates from best to worst based on overall quality. Consider how well each summary performs across all criteria to determine the final ranking.

^a https://en.wikipedia.org/wiki/C2_Proficiency

Figure D.8: A snapshot of the experimental instructions seen by our human raters.

E

Appendix for Chapter VII

E.1 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, and F1 are defined as:

$$R_L = \frac{LCS(\hat{y}_i, y_i)}{|y_i|} \quad (\text{E.1})$$

$$P_L = \frac{LCS(\hat{y}_i, y_i)}{|\hat{y}_i|} \quad (\text{E.2})$$

$$F_L = \frac{(1 + \beta^2) \cdot R_L \cdot P_L}{R_L + \beta^2 \cdot P_L} \quad (\text{E.3})$$

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 $\mathcal{Y}_i = \{y_i^{(1)}, y_i^{(2)}, \dots, y_i^{(k_i)}\}$ for a query and a generated answer \hat{y}_i , the instance-level DR score is defined as:

$$\text{DR}_i = \frac{1}{|\mathcal{Y}_i|} \sum_{j=1}^{|\mathcal{Y}_i|} \mathbf{1}[y_i^{(j)} \subseteq \hat{y}_i] \quad (\text{E.4})$$

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:

$$\text{DR} = \frac{1}{N} \sum_{i=1}^N \text{DR}_i. \quad (\text{E.5})$$

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. 2020b), 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 $\mathcal{Y}_i = \{y_i^{(1)}, \dots, y_i^{(k_i)}\}$, SARI is defined as:

$$\text{SARI} = \frac{1}{3} \left(\text{Add}_{F_1} + \text{Keep}_{F_1} + \text{Del}_{F_1} \right) \quad (\text{E.6})$$

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.

E.2 Details of Baselines

Here we describe the baselines used for comparison:

- **Standard RAG** (Lewis et al. 2020b) 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. 2023b) 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. 2024b) 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. 2026) 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. 2025c) 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.

E.3 Prompts for Disco-RAG

Appendix Figure E.3, Figure E.4, Figure E.5 and Figure E.11 present the prompts used in Disco-RAG.

Relation Definitions for Intra-chunk RST Tree Construction
<p>Relation Definitions:</p> <ul style="list-style-type: none"> - 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 E.1: Relation definitions for intra-chunk RST tree construction.

Relation Definitions for Inter-chunk Rhetorical Graph Construction
<p>Relation Definitions:</p> <ul style="list-style-type: none"> - 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 E.2: Relation definitions for inter-chunk rhetorical graph construction.

E.4 Human Evaluation Guidelines

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 units (EDUs).
2. Determine the most important EDU (this becomes the root nucleus).
3. For each other EDU, decide: Is it a nucleus (core) or a 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.

TEXT TO ANALYZE: {chunk_i}

Figure E.3: Prompt for intra-chunk RST tree construction. The relation definitions are provided in [Figure E.1](#).

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 E.4: Prompt for listwise discourse relation inference. The relation definitions are provided in [Figure E.2](#).

Prompt for Discourse-Driven Planning
<p>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.</p> <p>Inputs:</p> <ol style="list-style-type: none"> 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. <p>Critical instructions:</p> <ol style="list-style-type: none"> 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. <p>Required output format:</p> <p>PLAN: <one paragraph in natural language that describes the planned organization of the answer></p> <p>TEXT TO ANALYZE: {query, chunks, RST trees, rhetorical graph}</p>

Figure E.5: Prompt for discourse-driven planning.

Prompt for Full Context Generation
<p>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.</p> <p>Inputs:</p> <ol style="list-style-type: none"> 1. The user query. 2. The full document. <p>Critical instructions:</p> <ol style="list-style-type: none"> 1. The answer must directly address the user’s 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. <p>Required output format:</p> <p>ANSWER <one paragraph or multiple paragraphs in natural language></p> <p>TEXT TO ANALYZE {query, document}</p>

Figure E.6: Prompt for full context generation used in our baseline.

Prompt for Standard RAG
<p>You are an expert in retrieval-augmented generation. Your task is to answer the user query using only the retrieved text chunks as evidence.</p> <p>Inputs:</p> <ol style="list-style-type: none"> 1. The user query. 2. Retrieved text chunks. <p>Critical instructions:</p> <ol style="list-style-type: none"> 1. The answer must directly address the user’s 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. <p>Required output format:</p> <p>ANSWER <one paragraph or multiple paragraphs in natural language> TEXT TO ANALYZE {query, chunks}</p>

Figure E.7: Prompt for standard RAG used in our baseline.

Prompt for Retrieve-and-Plan Baseline
<p>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.</p> <p>Inputs:</p> <ol style="list-style-type: none"> 1. The user query. 2. Retrieved text chunks. <p>Critical instructions:</p> <ol style="list-style-type: none"> 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. <p>Required output format:</p> <p>PLAN <one paragraph plan> ANSWER <one paragraph or multiple paragraphs in natural language> TEXT TO ANALYZE {query, chunks}</p>

Figure E.8: Prompt for the retrieve-and-plan baseline used in our ablation study.

Prompt for Plan-and-Retrieve Baseline
<p>You are an expert in retrieval-augmented generation. Your task is to support a plan-guided retrieval procedure and then answer the user query.</p> <p>Stage 1 Given only the user query, write a short plan and a retrieval hint that summarizes what evidence should be retrieved.</p> <p>Stage 2 After retrieving all text chunks using the retrieval hint, write the final answer using only the retrieved text chunks as evidence.</p> <p>Inputs:</p> <ol style="list-style-type: none"> 1. The user query. 2. Retrieved text chunks returned after plan-guided retrieval. <p>Critical instructions:</p> <ol style="list-style-type: none"> 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 help 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. <p>Required output format:</p> <p>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}</p>

Figure E.9: 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 E.10: Prompt for discourse marker inference used in the shallow discourse marker base-line.

Prompt for Discourse-Guided RAG
<p>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:</p> <p>Inputs:</p> <ol style="list-style-type: none"> 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. <p>Critical instructions:</p> <ol style="list-style-type: none"> 1. The answer must directly address the user's 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. <p>Required output format: ANSWER: <a single coherent paragraph or multi-paragraph answer grounded in discourse structures></p> <p>Validation requirements:</p> <ul style="list-style-type: none"> - 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. <p>TEXT TO ANALYZE: {query, chunks, RST trees, rhetorical graph, discourse-aware plan}</p>

Figure E.11: 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 E.12: Guidelines presented to human raters for the SciNews dataset evaluation.

List of Figures

1.1	Dissertation roadmap showing how discourse signals are incorporated across the five research questions.	11
2.1	Evidence-conditioned generation under limited evidence access.	19
2.2	Comparison of two text generation paradigms.	20
2.3	Comparison of different attention patterns.	21
2.4	An example of RST tree: [<i>Utilizing discourse structure to enhance text summarization is beneficial.</i>] ^{EDU1} [<i>This technique can be used to identify key ideas and capture often overlooked nuances.</i>] ^{EDU2} [<i>Accurate capture of these complex structures facilitates the generation of good summaries.</i>] ^{EDU3}	23
2.5	An illustration of the Question under Discussion (QUD) framework.	25
2.6	Overview of parameter-efficient adaptation.	27
2.7	Illustration of low-rank adaptation and quantization-aware variants.	28
2.8	Standard retrieval-augmented generation pipeline. An external corpus is segmented into chunks, encoded into dense vectors, and indexed. At query time, the top- k chunks are retrieved based on embedding similarity and concatenated as context for the language model.	30
2.9	Structural limitations of standard RAG pipelines. Retrieved chunks are concatenated without explicit rhetorical connections. When chunks contain conflicting or conditionally scoped claims, the generator lacks the structural signals needed to reconcile them, which can lead to overgeneralization or factual inconsistency.	31
2.10	Planning-based generation in which a plan b is generated from evidence and then used to condition the final output.	33
2.11	Planning as an interface for controllability and verification.	34

2.12	Planning as a mechanism for linking multimodal evidence to sentence-level intents.	35
3.1	An example of RST tree: [<i>Rhetorical structure theory (RST) is a theory of text organization.</i>] ^{EDU1} [<i>Although the RST structure is difficult to annotate,</i>] ^{EDU2} [<i>there are still many scholars who have studied it.</i>] ^{EDU3}	43
3.2	Labeled discourse distributions.	45
3.3	Model architecture. We show a schematic diagram of incorporating LDD tensor into the attention layer of the model. Specifically, X is the text embedding matrix, and LDD is incorporated with the attention matrix S in the form of element-wise multiplication. In order to ensure the consistency of matrix shape, we also apply an identical <i>chunk</i> method as Longformer in LDD.	46
3.4	Sentence alignment distributions. $L = \text{Longformer}$, $R(\text{w/o}) = \text{RSTformer}(\text{w/o relations})$, $R(\text{w}) = \text{RSTformer}(\text{w/ relations})$	53
3.5	N-gram novelty. $L = \text{Longformer}$, $R(\text{w/o}) = \text{RSTformer}(\text{w/o relations})$, $R(\text{w}) = \text{RSTformer}(\text{w/ relations})$, $BC = \text{Booksum Chapter}$, $ML = \text{Multi-LexSum}$	54
3.6	Consistency check. $L = \text{Longformer}$, $R(\text{w/o}) = \text{RSTformer}(\text{w/o relations})$, $R(\text{w}) = \text{RSTformer}(\text{w/ relations})$, $BC = \text{Booksum Chapter}$, $ML = \text{Multi-LexSum}$	54
4.1	RST distribution.	62
4.2	Model architecture. The diagram illustrates the integration of the RST matrix into the LoRA model. The left side is the original LoRA, while the right side depicts our proposed method RST-LoRA.	65
4.3	Factual consistency analysis.	71
4.4	Impact of different r values on three datasets.	72
5.1	An example from VISTA: a conference presentation video (left) paired with the abstract of the corresponding paper (right). This data sample (Mallen et al. 2023) was presented at ACL 2023 and received the Best Video Recordings award.	84
5.2	Venue distribution of the VISTA dataset.	86

5.3	Distribution of summary sentences, summary tokens, video durations, and video shots in VISTA.	87
5.4	GPT-o1 generates plans based on reference summaries. Each question q_i corresponds to a summary sentence t_i , which we assume constitutes its answer. Index i ranges from 1 to the number of summary sentences.	89
5.5	Noise in plan generation impacts summarization performance. FRR is a shorthand for Full Random Replacement, and RR for Random Replacement. RAST is a SOTA question generation method.	95
5.6	Impact of text context for plan generation.	97
5.7	Human evaluation results. Human-written summaries consistently outperform all neural models.	100
5.8	LMM-as-a-Judge evaluation results showing that human-written summaries consistently outperform neural models.	102
6.1	An excerpt of a source document paired with its summary (bottom); the explanatory sentence and its target are highlighted in green and orange, respectively. The RST tree (top) for the text corresponding to the two sentences shows they are linked by the rhetorical relation <i>Elaboration</i>	113
6.2	We use DMRST (Liu et al. 2021e) to extract explanatory (e) EDUs and their target (t) EDUs from reference summaries, and then feed this data into GPT-4o to generate plans (b).	115
6.3	Summary quality as a function of different RST parsers.	124
6.4	Summary quality as a function of different question generation methods.	125
6.5	Human and GPT-4o-based evaluation results across datasets and evaluation dimensions.	130
7.1	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 <i>Satellite</i> (the supplementary part), and N denotes <i>Nucleus</i> (the core part).	141
7.2	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).	143

7.3	Performance comparison under varying chunk size (a), Top- k value (b), and retrieval noise level (c).	151
7.4	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.	152
7.5	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.	158
7.6	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.	159
7.7	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.	160
B.1	Prompt for zero-shot summaries generation.	179
B.2	Prompt for in-context summaries generation.	179
B.3	Prompt for summaries evaluation.	180
B.4	A snapshot of the experimental instructions provided to human evaluators.	181
C.1	A random sample from the VISTA dataset, originating from Fernandes et al. (2023).	184
C.2	Quality control guidelines.	185
C.3	Prompt to generate summaries without plans.	185
C.4	Prompt for question generation.	185
C.5	Prompt for PG model.	185
C.6	Prompt for SG model.	186
C.7	Prompt used by GPT-o1 to generate irrelevant questions.	186
C.8	Summary readability modification.	186
C.9	Summary length modification.	186
C.10	Plan readability modification.	186

C.11	A snapshot of the experimental instructions provided to human evaluators. . .	187
C.12	Prompt for GPT-o1 to evaluate summary quality.	188
D.1	Prompt used by GPT-4o and Mistral to generate plan questions based on target sentences and their context.	192
D.2	Prompt used by Mistral _{ZS} , GPT-4o _{ZS} and Mistral _{FT} to generate summaries.	192
D.3	Prompt used by Mistral _{ICL} and GPT-4o _{ICL} to generate summaries.	193
D.4	Prompt used by GPT-4o to identify named entities for Blueprint model. . .	193
D.5	Prompt used by Plan-Input and Plan-Output models.	193
D.6	Prompt used by GPT-4o and Mistral to directly extract explanatory content. .	194
D.7	Prompt used by GPT-4o to generate irrelevant questions.	194
D.8	A snapshot of the experimental instructions seen by our human raters. . . .	195
E.1	Relation definitions for intra-chunk RST tree construction.	201
E.2	Relation definitions for inter-chunk rhetorical graph construction.	202
E.3	Prompt for intra-chunk RST tree construction. The relation definitions are provided in Figure E.1.	204
E.4	Prompt for listwise discourse relation inference. The relation definitions are provided in Figure E.2.	205
E.5	Prompt for discourse-driven planning.	206
E.6	Prompt for full context generation used in our baseline.	206
E.7	Prompt for standard RAG used in our baseline.	207
E.8	Prompt for the retrieve-and-plan baseline used in our ablation study.	207
E.9	Prompt for the plan-and-retrieve baseline used in our ablation study.	208
E.10	Prompt for discourse marker inference used in the shallow discourse marker baseline.	209
E.11	Prompt for discourse-guided RAG.	210
E.12	Guidelines presented to human raters for the SciNews dataset evaluation. . .	211

List of Tables

3.1	RST relation category.	48
3.2	Datasets statistics.	49
3.3	Model performance. The bold numbers represent the best results with respect to the given test set. Δ and Φ represent the improvement of our model compared to the baseline and SOTA models, respectively. \dagger and \ddagger indicate statistical significance ($p < 0.05$) against the baseline model via T-test and Kolmogorov-Smirnov test. Each result of the three distinct SOTA models is directly replicated from their original papers.	51
3.4	F1 scores for ablation study.	52
3.5	Human evaluation results.	53
4.1	Performance of different RST variants.	68
4.2	Model performance. The bold numbers represent the best results concerning the given test set. \dagger and \ddagger indicate statistical significance ($p < 0.05$) of our final model (RST _w ^p -LoRA) against the FFT and LoRA model via paired t-test based on the same backbone, respectively. FFT for full fine-tuning, ZS for zero-shot learning, and ICL for in-context learning. Each result of the SOTA models is directly replicated from their original papers.	69
4.3	F1 scores for ablation study.	70
4.4	Impact of random masking on the parser.	73
4.5	Human evaluation results.	74
4.6	GPT-4 evaluation results on BC dataset.	74
4.7	GPT-4 evaluation results on ML dataset.	75
4.8	GPT-4 evaluation results on eLife dataset.	75
4.9	Comparison of target summary with RST-LoRA and GPT-4 outputs.	76
4.10	Comparison of target summary with RST-LoRA and GPT-4 outputs.	77

5.1	Comprison of video-to-text summarization datasets. #Videos = the number of videos, whereas VideoLen and SumLen refer to the average of video duration (in minutes) and the average number of summary tokens.	87
5.2	Key statistics of the VISTA dataset, showcasing the average video length and shot count, summary characteristics (sentence and token counts), syntactic complexity (dependency tree depth), and lexical diversity (Type-Token Ratio and Distinct n-gram scores).	88
5.3	Model version details.	91
5.4	Model performance on VISTA dataset. In Plan-mPlug-0w13*, only the PG module is trained. Plans generated by the PG on the test set serve as input to the SG module for zero-shot inference (no training is applied to the SG module). Symbols † and ‡ indicate that the performance of Plan-mPlug-0w13 is significantly ($p < 0.05$) different from LLaVA-NeXT-Interleave (third best) and mPLUG-0w13 (second best), when using the paired t-test.	93
5.5	Performance comparison of different modality combinations.	94
5.6	Performance comparison of different plan generation ablations under full fine-tuning settings.	95
5.7	Model performance under different video context configurations (full fine-tuning). The video content at the end is more helpful for summary generation.	96
5.8	Control experiment for summary readability. FRE = Flesch Reading Ease.	98
5.9	Control experiment for summary length.	98
5.10	Performance of baseline vs. planning models in non-video settings across different learning regimes. Each cell shows the result <i>before</i> → <i>after</i> applying the planning method.	99
5.11	Comparative case study of model outputs. Hallucinations are highlighted in red.	105
5.12	Decreasing reading difficulty: plan modifications vs instruction adjustments. Hallucinations are in red.	106
5.13	Increasing reading difficulty: plan modifications vs. instruction adjustments. Hallucinations are in red.	106
5.14	Deleting 30% of plan questions vs. instruction adjustments. Hallucinations are highlighted in red.	107
6.1	RST explanatory relations.	116

6.2	Model performance on three lay summarization datasets. Bold numbers represent the best results achieved by the models in each test set, excluding human results from the comparison. <code>ExpRatio</code> is the number of explanatory EDUs over all EDUs in summaries. Symbols [†] and [‡] denote that Plan-Input is statistically significant ($p < 0.05$) against <code>Mistral_{FT}</code> and <code>Blueprint_{MT}</code> using paired t-test, respectively.	121
6.3	Comparison of different plan generation strategies using Plan-Input and Mistral in fine-tuned and zero-shot settings.	123
6.4	Proportion of explanatory EDUs in predicted summaries for models with (De1) and without (NoDe1) deletions in the plan. In the NoDe1 setting, the model regulates the proportion of explanatory content in the plan and summary. In the De1 setting, the plan is manipulated by removing specific explanations. $\Delta\text{RLsum} = \text{RLsum before plan deletion} - \text{RLsum after plan deletion}$. 127	127
6.5	Case study 1: we use red font to mark hallucinated information. This sample comes from the test split of the eLife dataset (Goldsack et al. 2022).	133
6.6	Case study 2: we use red font to mark hallucinated information. This sample comes from the test split of the eLife dataset (Goldsack et al. 2022).	134
7.1	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. (2025c). We use bold red to indicate the best results and <u>blue underlined text</u> to indicate the second-best results.	147
7.2	Performance on the ASQA benchmark. Disco-RAG consistently outperforms standard RAG across all metrics. It also surpasses existing SOTA methods on most dimensions.	148
7.3	Performance on the SciNews dataset. Disco-RAG beats both zero-shot and standard RAG, and often surpasses prior SOTA across multiple metrics. . .	149

7.4	Ablation study of the three modules in Disco-RAG with Llama-3.3-70B. <i>w/o RST tree</i> removes intra-chunk modeling, <i>w/o rhetorical graph</i> removes inter-chunk modeling, and <i>w/o planning</i> removes discourse-aware planning. We additionally report two generic planning baselines built on standard RAG. <i>retrieve-and-plan</i> generates a free-form plan conditioned on retrieved chunks before generation, and <i>plan-and-retrieve</i> first generates a free-form plan from the query and then performs a retrieval step guided by this plan. . . .	150
7.5	Evaluation of the RST parser on the RST-DT benchmark following the protocol of Maekawa et al. (2024).	153
7.6	Comparison of standard RAG, a shallow discourse marker variant, and Disco-RAG on the Loong benchmark with Llama-3.3-70B.	154
7.7	Mixed-model results on the Loong benchmark. <i>Structurer</i> refers to the model used for RST parsing, rhetorical graph construction, and planning; <i>Generator</i> refers to the model used for final answer generation. 8B = Llama-3.1-8B; 70B = Llama-3.3-70B.	155
7.8	Impact of supervised fine-tuning (SFT) and discourse conditioning.	156
7.9	Average human ratings on SciNews. Scores are computed on a three-point Likert scale, and higher values indicate better performance.	157
D.1	Words and phrases used to heuristically identify discourse relations signaling explanations.	191

Bibliography

Griffin Adams, Alex Fabbri, Faisal Ladhak, Noémie Elhadad, and Kathleen McKeown. 2023. **Generating EDU Extracts for Plan-Guided Summary Re-Ranking**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 2680–2697.

Rilwan Adewoyin, Ritabrata Dutta, and Yulan He. 2022. **RSTGen: Imbuing Fine-Grained Interpretable Control into Long-Form Text Generators**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 1822–1835.

Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. 2024. **Tuning Large Multimodal Models for Videos using Reinforcement Learning from AI Feedback**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 923–940.

Mohammad AkbarTajari, Sara Rajaei, and Mohammad Taher Pilehvar. 2022. **An Empirical Study on the Transferability of Transformer Modules in Parameter-efficient Fine-tuning**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 10617–10625.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. **Unsupervised opinion summarization with content planning**. In *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 14, pp. 12489–12497.

Anthropic. 2024. *Claude 3.5 - Sonnet*. <https://www.anthropic.com/news/claude-3-5-sonnet>. Accessed: 2024-12-06.

Dawit Mureja Argaw, Seunghyun Yoon, Fabian Caba Heilbron, Hanieh Deilamsalehy, Trung Bui, Zhaowen Wang, Franck Deroncourt, and Joon Son Chung. 2024. *Scaling Up Video Summarization Pretraining with Large Language Models*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8332–8341.

Akari Asai, Mohammadreza Salehi, Matthew Peters, and Hannaneh Hajishirzi. 2022. *ATTEMPT: Parameter-Efficient Multi-task Tuning via Attentional Mixtures of Soft Prompts*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 6655–6672.

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*.

Katherine Atwell, Anthony Sicilia, Seong Jae Hwang, and Malihe Alikhani. 2022. *The Change that Matters in Discourse Parsing: Estimating the Impact of Domain Shift on Parser Error*. In *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 824–845.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. *Generating Scientific Definitions with Controllable Complexity*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 8298–8317.

Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. *Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing*. In *ACM Trans. Comput. Hum. Interact.* 30.5, 74:1–74:38.

Kartikeya Badola, Shachi Dave, and Partha Talukdar. 2023. *Parameter-Efficient Fine-tuning for Robust Continual Multilingual Learning*. In *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 9763–9780.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. *Hallucination of Multimodal Large Language Models: A Survey*. In *CoRR* abs/2404.18930.

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. *Correcting Diverse Factual Errors in Abstractive Summarization via Post-Editing and Language Model Infilling*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 9818–9830.

Vidhisha Balachandran, Artidoro Pagnoni, Jay Yoon Lee, Dheeraj Rajagopal, Jaime Carbonell, and Yulia Tsvetkov. 2021. *StructSum: Summarization via Structured Representations*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 2575–2585.

Satanjeev Banerjee and Alon Lavie. 2005. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72.

David I Beaver, Craige Roberts, Mandy Simons, and Judith Tonhauser. 2017. *Questions under discussion: Where information structure meets projective content*. In *Annual Review of Linguistics* 3.1, pp. 265–284.

Yonatan Belinkov and Yonatan Bisk. 2018. *Synthetic and Natural Noise Both Break Neural Machine Translation*. In *International Conference on Learning Representations*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. *Longformer: The Long-Document Transformer*. In *CoRR* abs/2004.05150.

Anton Benz and Katja Jasinskaja. 2017. *Questions Under Discussion: From Sentence to Discourse*. In *Discourse Processes* 54.3, pp. 177–186.

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*. Lisbon, Portugal: Association for Computational Linguistics, pp. 2212–2218.

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*.

Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. 2023. **SIMSUM: Document-level Text Simplification via Simultaneous Summarization**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 9927–9944.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. **Language Models are Few-Shot Learners**. In *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 1877–1901.

Emanuele Bugliarelo and Naoaki Okazaki. 2020. **Enhancing Machine Translation with Dependency-Aware Self-Attention**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1618–1627.

Isabel Cachola, Daniel Khashabi, and Mark Dredze. 2025. **Evaluating the Evaluators: Are readability metrics good measures of readability?** In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, pp. 24011–24027.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. **TLDR: Extreme Summarization of Scientific Documents**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 4766–4777.

Gerard Canal, Senka Krivic, Paul Luff, and Andrew Coles. 2022. **PlanVerb: Domain-Independent Verbalization and Summary of Task Plans**. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational*

Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. AAAI Press, pp. 9698–9706.

Meng Cao, Yue Dong, and Jackie Cheung. 2022a. **Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 3340–3354.

Shuyang Cao and Lu Wang. 2022b. **HIBRIDS: Attention with Hierarchical Biases for Structure-aware Long Document Summarization**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 786–807.

Shuyang Cao and Lu Wang. 2024. **AWESOME: GPU Memory-constrained Long Document Summarization using Memory Mechanism and Global Salient Content**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico: Association for Computational Linguistics, pp. 5925–5941.

Ronald Cardenas, Bingsheng Yao, Dakuo Wang, and Yufang Hou. 2023. **‘Don’t Get Too Technical with Me’: A Discourse Structure-Based Framework for Automatic Science Journalism**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 1186–1202.

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*.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. **Overview and Insights from the Shared Tasks at Scholarly Document Processing 2020: CL-SciSumm, LaySumm and LongSumm**. In *Proceedings of the First Workshop on Scholarly Document Processing*. Online: Association for Computational Linguistics, pp. 214–224.

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). Vienna, Austria: Association for Computational Linguistics, pp. 2607–2622.

Brian Chen, Xiangyuan Zhao, and Yingnan Zhu. 2024a. **Personalized Video Summarization by Multimodal Video Understanding**. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. CIKM '24. Boise, ID, USA: Association for Computing Machinery, pp. 4382–4389.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024b. **MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark**. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. Vol. 235. Proceedings of Machine Learning Research. PMLR / OpenReview.net, pp. 6562–6595.

Guanzheng Chen, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang. 2022. **Revisiting Parameter-Efficient Tuning: Are We Really There Yet?** In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 2612–2626.

Huadong Chen, Shujian Huang, David Chiang, and Jiajun Chen. 2017. **Improved Neural Machine Translation with a Syntax-Aware Encoder and Decoder**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1936–1945.

Jiaao Chen and Diyi Yang. 2021. **Structure-Aware Abstractive Conversation Summarization via Discourse and Action Graphs**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 1380–1391.

Jiaao Chen, Aston Zhang, Xingjian Shi, Mu Li, Alex Smola, and Diyi Yang. 2023. **Parameter-Efficient Fine-Tuning Design Spaces**. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Siyuan Chen, Hang Ding, Kangxiaoyu, and Jiechao Gao. 2026. **GRO-RAG: Gradient-aware Re-rank Optimization for Multi-source Retrieval-Augmented Generation**. In *The Fourteenth International Conference on Learning Representations*.

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. 2024c. **M³AV: A Multimodal, Multigenre, and Multipurpose Audio-Visual Academic Lecture Dataset**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 9041–9060.

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)*. Vienna, Austria: Association for Computational Linguistics, pp. 31877–31899.

Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. 2023. **Can We Edit Multimodal Large Language Models?** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 13877–13888.

Elena Chistova. 2023. **End-to-End Argument Mining over Varying Rhetorical Structures**. In *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 3376–3391.

Elena Chistova. 2024. **Bilingual Rhetorical Structure Parsing with Large Parallel Annotations**. In *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, pp. 9689–9706.

Sangwoo Cho, Franck Dernoncourt, Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, and Fei Liu. 2021. **StreamHover: Livestream Transcript Summarization and Annotation**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 6457–6474.

Krzysztof Marcin Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. 2021.

Rethinking Attention with Performers. In *International Conference on Learning Representations*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. *Qwen2-Audio Technical Report*. In *CoRR* abs/2407.10759.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. **A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents.** In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 615–621.

Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. **Automatic Generation of Natural Language Explanations.** In *Companion Proceedings of the 23rd International Conference on Intelligent User Interfaces*. IUI '18 Companion. Tokyo, Japan: Association for Computing Machinery.

Aldan Creo, Manuel Lama, and Juan Carlos Vidal. 2023. *Prompting LLMs with content plans to enhance the summarization of scientific articles.* In *CoRR* abs/2312.08282.

Noam Dahan and Gabriel Stanovsky. 2025. **The State and Fate of Summarization Datasets: A Survey.** 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)*. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 7259–7278.

Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Re. 2022. **FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness.** In *Advances in Neural Information Processing Systems*.

Forrest Davis and Marten van Schijndel. 2020. **Discourse structure interacts with reference but not syntax in neural language models.** In *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, pp. 396–407.

Kordula De Kuthy, Madeeswaran Kannan, Haemanth Santhi Ponnusamy, and Detmar Meurers. 2020. **Towards automatically generating Questions under Discussion to link**

information and discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 5786–5798.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. **QLoRA: Efficient Finetuning of Quantized LLMs**. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hang Ding, Yilun Zhao, Tiansheng Hu, Manasi Patwardhan, and Arman Cohan. 2026. **SciRAG: Adaptive, Citation-Aware, and Outline-Guided Retrieval and Synthesis for Scientific Literature**. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Rabat, Morocco: Association for Computational Linguistics, pp. 6440–6460.

Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. 2023. **Sparse Low-rank Adaptation of Pre-trained Language Models**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 4133–4145.

Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. **Discourse-Aware Unsupervised Summarization for Long Scientific Documents**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 1089–1102.

zican Dong, Junyi Li, Xin Men, Xin Zhao, Bingning Wang, Zhen Tian, weipeng chen, and Ji-Rong Wen. 2024. **Exploring Context Window of Large Language Models via Decomposed Positional Vectors**. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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, and Jonathan Larson. 2024. **From Local to Global: A Graph RAG Approach to Query-Focused Summarization**. In *CoRR* abs/2404.16130.

Florian Eichin, Yang Janet Liu, Barbara Plank, and Michael A. Hedderich. 2025. **Probing LLMs for Multilingual Discourse Generalization Through a Unified Label Set**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, pp. 18665–18684.

Mohamed Elaraby and Diane Litman. 2022. **ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining**. In *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 6187–6194.

Allyson Ettinger. 2020. *What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models*. In *Transactions of the Association for Computational Linguistics* 8, pp. 34–48.

Mehwish Fatima and Michael Strube. 2023. **Cross-lingual Science Journalism: Select, Simplify and Rewrite Summaries for Non-expert Readers**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 1843–1861.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. **When Does Translation Require Context? A Data-driven, Multilingual Exploration**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 606–626.

Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2024. **Qlarify: Recursively Expandable Abstracts for Dynamic Information Retrieval over Scientific Papers**. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. UIST '24. Pittsburgh, PA, USA: Association for Computing Machinery.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. **BLINK: Multimodal Large Language Models Can See but Not Perceive**. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIII*. Vol. 15081. Lecture Notes in Computer Science. Springer, pp. 148–166.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2021. **MM-AVS: A Full-Scale Dataset for Multi-modal Summarization**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 5922–5926.

Yingxue Fu. 2025. **A Survey of QUD Models for Discourse Processing**. 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)*. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 1722–1732.

Yingxue Fu, Mark-Jan Nederhof, and Anais Ollagnier. 2025. **A Topicality-Driven QUD Model for Discourse Processing**. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Avignon, France: Association for Computational Linguistics, pp. 214–230.

Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Kyle Lo, Asli Celikyilmaz, and Yejin Choi. 2021. **Discourse Understanding and Factual Consistency in Abstractive Summarization**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 435–447.

Shen Gao, Xin Cheng, Mingzhe Li, Xiuying Chen, Jinpeng Li, Dongyan Zhao, and Rui Yan. 2023. **Dialogue Summarization with Static-Dynamic Structure Fusion Graph**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 13858–13873.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2025. **How to Train Long-Context Language Models (Effectively)**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, pp. 7376–7399.

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)*. Mexico City, Mexico: Association for Computational Linguistics, pp. 306–315.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bitá Nejat. 2014. **Abstractive Summarization of Product Reviews Using Discourse Structure**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1602–1613.

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. **DisCoDisCo at the DISRPT2021 Shared Task: A System for Discourse Segmentation, Classification, and Connective Detection**. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 51–62.

Marjan Ghazvininejad, Vladimir Karpukhin, Vera Gor, and Asli Celikyilmaz. 2022. **Discourse-Aware Soft Prompting for Text Generation**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 4570–4589.

Mozhdeh Gheini, Xuezhe Ma, and Jonathan May. 2023. **Know Where You’re Going: Meta-Learning for Parameter-Efficient Fine-Tuning**. In *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 11602–11612.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. **Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 10589–10604.

Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton, and Chenghua Lin. 2023. **Enhancing Biomedical Lay Summarisation with External Knowledge Graphs**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 8016–8032.

Qi Gou, Zehua Xia, Bowen Yu, Haiyang Yu, Fei Huang, Yongbin Li, and Nguyen Cam-Tu. 2023. **Diversify Question Generation with Retrieval-Augmented Style Transfer**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 1677–1690.

Naman Goyal and Jacob Eisenstein. 2016. **A Joint Model of Rhetorical Discourse Structure and Summarization**. In *Proceedings of the Workshop on Structured Prediction for NLP*. Austin, TX: Association for Computational Linguistics, pp. 25–34.

Naibin Gu, Peng Fu, Xiyu Liu, Zhengxiao Liu, Zheng Lin, and Weiping Wang. 2023. **A Gradient Control Method for Backdoor Attacks on Parameter-Efficient Tuning**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 3508–3520.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. **PPT: Pre-trained Prompt Tuning for Few-shot Learning**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 8410–8423.

Bin Guo, Hao Wang, Yasan Ding, Wei Wu, Shaoyang Hao, Yueqi Sun, and Zhiwen Yu. 2021a. **Conditional Text Generation for Harmonious Human-Machine Interaction**. In *ACM Trans. Intell. Syst. Technol.* 12.2.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontañón, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. **LongT5: Efficient Text-To-Text Transformer for Long Sequences**. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, pp. 724–736.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021b. **Automated Lay Language Summarization of Biomedical Scientific Reviews**. In *Proceedings of the AAAI Conference on Artificial Intelligence* 35.1, pp. 160–168.

Omkar Gurjar, Agam Goyal, and Eshwar Chandrasekharan. 2025. **ArgCMV: An Argument Summarization Benchmark for the LLM-era**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, pp. 21870–21883.

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*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org.

Grigorii Guz and Giuseppe Carenini. 2020. **Coreference for Discourse Parsing: A Neural Approach**. In *Proceedings of the First Workshop on Computational Approaches to Discourse*. Online: Association for Computational Linguistics, pp. 160–167.

HyoJung Han, Jordan Boyd-Graber, and Marine Carpuat. 2023. **Bridging Background Knowledge Gaps in Translation with Automatic Explicitation**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 9718–9735.

Kelvin Han and Claire Gardent. 2025a. **Generating Questions Under Discussion with Reinforcement Learning using Ranking and Scoring for Reward and Evaluation**. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. Mumbai, India: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, pp. 589–615.

Mingfei Han, Linjie Yang, Xiaojun Chang, Lina Yao, and Heng Wang. 2025b. **Shot2Story: A New Benchmark for Comprehensive Understanding of Multi-shot Videos**. In *The Thirteenth International Conference on Learning Representations*.

HangChen HangChen, Chao-Han Huck Yang, Jia-Chen Gu, Sabato Marco Siniscalchi, and Jun Du. 2025. **MISP-Meeting: A Real-World Dataset with Multimodal Cues for Long-form Meeting Transcription and Summarization**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, pp. 15479–15492.

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*. Los Angeles: Association for Computational Linguistics, pp. 128–136.

Bo He, Jun Wang, Jieli Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. **Align and Attend: Multimodal Summarization with Dual Contrastive Losses**. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, pp. 14867–14878.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. **Towards a Unified View of Parameter-Efficient Transfer Learning**. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyang Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhui Chen. 2024. **VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 2105–2123.

Freya Hewett, Hadi Asghari, and Manfred Stede. 2024. **Elaborative Simplification for German-Language Texts**. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Kyoto, Japan: Association for Computational Linguistics, pp. 29–39.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Larousilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. **Parameter-Efficient Transfer Learning for NLP**. In *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2790–2799.

Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2024. **mPLUG-PaperOwl: Scientific Diagram Analysis with the Multimodal Large Language Model**. In *ACM Multimedia 2024*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. **LoRA: Low-Rank Adaptation of Large Language Models**. In *International Conference on Learning Representations*.

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*. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 4145–4157.

Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2025. **V2Xum-LLM: cross-modal video summarization with temporal prompt instruction tuning**. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Yilun Hua, Nicholas Chernogor, Yuzhe Gu, Seoyeon Jeong, Miranda Luo, and Cristian Danescu-Niculescu-Mizil. 2024. **How did we get here? Summarizing conversation dynamics**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico: Association for Computational Linguistics, pp. 7452–7477.

Haoyu Huang, Yongfeng Huang, Yang Junjie, Zhenyu Pan, Yongqiang Chen, Kaili Ma, Hongzhi Chen, and James Cheng. 2025a. **Retrieval-Augmented Generation with Hierarchical Knowledge**. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Suzhou, China: Association for Computational Linguistics, pp. 6044–6060.

Kung-Hsiang Huang, Hou Pong Chan, May Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2025b. **From Pixels to Insights: A Survey on Automatic Chart Understanding in the Era of Large Foundation Models**. In *IEEE Transactions on Knowledge and Data Engineering* 37.5, pp. 2550–2568.

Yin Jou Huang and Sadao Kurohashi. 2021. **Extractive Summarization Considering Discourse and Coreference Relations based on Heterogeneous Graph**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 3046–3052.

Patrick Huber and Giuseppe Carenini. 2022. **Towards Understanding Large-Scale Discourse Structures in Pre-Trained and Fine-Tuned Language Models**. In *Proceedings of*

the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics, pp. 2376–2394.

Fantine Huot, Joshua Maynez, Chris Alberti, Reinald Kim Amplayo, Priyanka Agrawal, Constanza Fierro, Shashi Narayan, and Mirella Lapata. 2024. **μ PLAN: Summarizing using a Content Plan as Cross-Lingual Bridge**. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. St. Julian's, Malta: Association for Computational Linguistics, pp. 2146–2163.

Fantine Huot, Joshua Maynez, Shashi Narayan, Reinald Kim Amplayo, Kuzman Ganchev, Annie Priyadarshini Louis, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. **Text-Blueprint: An Interactive Platform for Plan-based Conditional Generation**. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 105–116.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024a. **Video ReCap: Recursive Captioning of Hour-Long Videos**. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, pp. 18198–18208.

Shayekh Bin Islam, Md Asib Rahman, K S M Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. 2024b. **Open-RAG: Enhanced Retrieval Augmented Reasoning with Open-Source Large Language Models**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, pp. 14231–14244.

Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2019. **Unsupervised Neural Single-Document Summarization of Reviews via Learning Latent Discourse Structure and its Ranking**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2142–2152.

Yeonseok Jeong, Minsoo Kim, Seung-won Hwang, and Byung-Hak Kim. 2025. **Agent-as-Judge for Factual Summarization of Long Narratives**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, pp. 23591–23608.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023a. *Mistral 7B*.

Pengcheng Jiang, Lang Cao, Ruike Zhu, Minhao Jiang, Yunyi Zhang, Jiaming Shen, Jimeng Sun, and Jiawei Han. 2026. **RAS: Retrieval-And-Structuring for Knowledge-Intensive LLM Generation**. In *The Fourteenth International Conference on Learning Representations*.

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

Sebastian Joseph, Lily Chen, Jan Trienes, Hannah G ke, Monika Coers, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2024. **FactPICO: Factuality Evaluation for Plain Language Summarization of Medical Evidence**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 8437–8464.

Jiaxin Ju, Ming Liu, Huan Yee Koh, Yuan Jin, Lan Du, and Shirui Pan. 2021. **Leveraging Information Bottleneck for Scientific Document Summarization**. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 4091–4098.

Yeonjoon Jung, Daehyun Ahn, Hyungjun Kim, Taesu Kim, and Eunhyeok Park. 2025. **GraLoRA: Granular Low-Rank Adaptation for Parameter-Efficient Fine-Tuning**. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. **Compacter: Efficient Low-Rank Hypercomplex Adapter Layers**. In *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 1022–1035.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. **Dense Passage Retrieval for Open-Domain Question Answering**. In *Proceedings of the 2020 Conference on Empirical Methods in Natu-*

ral Language Processing (EMNLP). Online: Association for Computational Linguistics, pp. 6769–6781.

Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2014. **Single Document Summarization based on Nested Tree Structure**. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 315–320.

Dongwhan Kim and Joonhwan Lee. 2019. **Designing an Algorithm-Driven Text Generation System for Personalized and Interactive News Reading**. In *Int. J. Hum. Comput. Interact.* 35.2, pp. 109–122.

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*. Singapore: Association for Computational Linguistics, pp. 996–1009.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. In.

Diederik P. Kingma and Jimmy Ba. 2015. **Adam: A Method for Stochastic Optimization**. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Wei-Jen Ko, Yating Wu, Cutter Dalton, Dananjay Srinivas, Greg Durrett, and Junyi Jessy Li. 2023. **Discourse Analysis via Questions and Answers: Parsing Dependency Structures of Questions Under Discussion**. In *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 11181–11195.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. **Top-Down RST Parsing Utilizing Granularity Levels in Documents**. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The*

Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press, pp. 8099–8106.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. **Grounding language models to images for multimodal inputs and outputs**. In *International Conference on Machine Learning*. PMLR, pp. 17283–17300.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021a. **Discourse Probing of Pretrained Language Models**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 3849–3864.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021b. **Top-down Discourse Parsing via Sequence Labelling**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 715–726.

Mateusz Krubiński and Pavel Pecina. 2023. **MLASK: Multimodal Summarization of Video-based News Articles**. In *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 910–924.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. **BOOKSUM: A Collection of Datasets for Long-form Narrative Summarization**. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 6536–6558.

Jonas M. Kübler, Wittawat Jitkrittum, Bernhard Schölkopf, and Krikamol Muandet. 2020. **Learning Kernel Tests Without Data Splitting**. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Keito Kudo, Haruki Nagasawa, Jun Suzuki, and Nobuyuki Shimizu. 2023. **A Challenging Multimodal Video Summary: Simultaneously Extracting and Generating Keyframe-Caption Pairs from Video**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 7380–7402.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. *SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization*. In *Transactions of the Association for Computational Linguistics* 10, pp. 163–177.

Neal Lawton, Anoop Kumar, Govind Thattai, Aram Galstyan, and Greg Ver Steeg. 2023. *Neural Architecture Search for Parameter-Efficient Fine-tuning of Large Pre-trained Language Models*. In *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 8506–8515.

Dosung Lee, Wonjun Oh, Boyoung Kim, Minyoung Kim, Joonsuk Park, and Paul Hong-suck 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)*. Vienna, Austria: Association for Computational Linguistics, pp. 341–359.

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)*. Vienna, Austria: Association for Computational Linguistics, pp. 879–893.

Tao Lei, Junwen Bai, Siddhartha Brahma, Joshua Ainslie, Kenton Lee, Yanqi Zhou, Nan Du, Vincent Y Zhao, Yuexin Wu, Bo Li, Yu Zhang, and Ming-Wei Chang. 2023. *Conditional Adapters: Parameter-efficient Transfer Learning with Fast Inference*. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. *The Power of Scale for Parameter-Efficient Prompt Tuning*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 3045–3059.

Guy Lev, Michal Shmueli-Scheuer, Jonathan Herzig, Achiya Jerbi, and David Konopnicki. 2019. *TalkSumm: A Dataset and Scalable Annotation Method for Scientific Paper Summarization Based on Conference Talks*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2125–2131.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. **BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024a. **SEED-Bench: Benchmarking Multimodal Large Language Models**. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, pp. 13299–13308.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. 2025a. **LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models**. In *The Thirteenth International Conference on Learning Representations*.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. **Multimodal Summarization for Asynchronous Collection of Text, Image, Audio and Video**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1092–1102.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. **A Diversity-Promoting Objective Function for Neural Conversation Models**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 110–119.

Jonathan Li, Will Aitken, Rohan Bhambhoria, and Xiaodan Zhu. 2023. **Prefix Propagation: Parameter-Efficient Tuning for Long Sequences**. In *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Toronto, Canada: Association for Computational Linguistics, pp. 1408–1419.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. 2020a. *Video Storytelling: Textual Summaries for Events*. In *IEEE Transactions on Multimedia* 22.2, pp. 554–565.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016b. **The Role of Discourse Units in Near-Extractive Summarization**. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Los Angeles: Association for Computational Linguistics, pp. 137–147.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. **Multimodal ArXiv: A Dataset for Improving Scientific Comprehension of Large Vision-Language Models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 14369–14387.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020b. **VMSMO: Learning to Generate Multimodal Summary for Video-based News Articles**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 9360–9369.

Mufei Li, Siqi Miao, and Pan Li. 2025b. **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*.

Xiang Lisa Li and Percy Liang. 2021. **Prefix-Tuning: Optimizing Continuous Prompts for Generation**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 4582–4597.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024c. **LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models**. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVI*. Milan, Italy: Springer-Verlag, pp. 323–340.

Yixiao Li, Yifan Yu, Chen Liang, Nikos Karampatziakis, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2024d. **LoftQ: LoRA-Fine-Tuning-aware Quantization for Large Language Models**. In *The Twelfth International Conference on Learning Representations*.

Zhenwen Li, Wenhao Wu, and Sujian Li. 2020c. **Composing Elementary Discourse Units in Abstractive Summarization**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6191–6196.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2025c. **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. Sydney NSW, Australia: Association for Computing Machinery, pp. 334–343.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024a. *Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions*. In *ACM Comput. Surv.* 56.10.

Yi Liang, You Wu, Honglei Zhuang, Li Chen, Jiaming Shen, Yiling Jia, Zhen Qin, Sumit Sanghai, Xuanhui Wang, Carl Yang, and Michael Bendersky. 2024b. **Integrating Planning into Single-Turn Long-Form Text Generation**. In *CoRR* abs/2410.06203.

Baohao Liao, Yan Meng, and Christof Monz. 2023. **Parameter-Efficient Fine-Tuning without Introducing New Latency**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 4242–4260.

Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024a. **Video-LLaVA: Learning United Visual Representation by Alignment Before Projection**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 5971–5984.

Chin-Yew Lin. 2004. **ROUGE: A Package for Automatic Evaluation of Summaries**. In *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81.

Chin-Yew Lin and Eduard Hovy. 2003. **Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics**. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 150–157.

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2024b. **VideoXum: Cross-Modal Visual and Textural Summarization of Videos**. In *IEEE Trans. Multim.* 26, pp. 5548–5560.

Teng Lin, Yizhang Zhu, Yuyu Luo, and Nan Tang. 2025. **SRAG: Structured Retrieval-Augmented Generation for Multi-Entity Question Answering over Wikipedia Graph**. In *CoRR* abs/2503.01346.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. **Open Sesame: Getting inside BERT’s Linguistic Knowledge**. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 241–253.

Danyang Liu, Mirella Lapata, and Frank Keller. 2023a. **Visual Storytelling with Question-Answer Plans**. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, pp. 5800–5813.

Dongqi Liu and Vera Demberg. 2023b. **ChatGPT vs Human-authored Text: Insights into Controllable Text Summarization and Sentence Style Transfer**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*. Toronto, Canada: Association for Computational Linguistics, pp. 1–18.

Dongqi Liu and Vera Demberg. 2024a. **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)*. Mexico City, Mexico: Association for Computational Linguistics, pp. 2200–2220.

Dongqi Liu, Xudong Hong, Pin-Jie Lin, Ernie Chang, and Vera Demberg. 2022a. **Two-Stage Movie Script Summarization: An Efficient Method For Low-Resource Long Document Summarization**. In *Proceedings of the Workshop on Automatic Summarization for Creative Writing*. Gyeongju, Republic of Korea: Association for Computational Linguistics, pp. 57–66.

Dongqi Liu and Khalil Sima'an. 2022b. **Passing Parser Uncertainty to the Transformer: Labeled Dependency Distributions for Neural Machine Translation**. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. Ghent, Belgium: European Association for Machine Translation, pp. 41–50.

Dongqi Liu, Yifan Wang, and Vera Demberg. 2023c. **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)*. Toronto, Canada: Association for Computational Linguistics, pp. 5574–5590.

Dongqi Liu, Yifan Wang, Jia Loy, and Vera Demberg. 2024b. **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)*. Torino, Italia: ELRA and ICCL, pp. 14429–14444.

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)*. Vienna, Austria: Association for Computational Linguistics, pp. 6187–6210.

Dongqi Liu, Xi Yu, Vera Demberg, and Mirella Lapata. 2025b. **Explanatory Summarization with Discourse-Driven Planning**. In *Transactions of the Association for Computational Linguistics* 13, pp. 1146–1170.

Hui Liu and Xiaojun Wan. 2021a. **Video Paragraph Captioning as a Text Summarization Task**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, pp. 55–60.

Hui Liu and Xiaojun Wan. 2023d. **Models See Hallucinations: Evaluating the Factuality in Video Captioning**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 11807–11823.

Ran Liu, Ming Liu, Min Yu, He Zhang, Jianguo Jiang, Gang Li, and Weiqing Huang. 2024c. **SumSurvey: An Abstractive Dataset of Scientific Survey Papers for Long Document Summarization**. In *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, pp. 9632–9651.

Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023e. **RECAP: Retrieval-Enhanced Context-Aware Prefix Encoder for Personalized Dialogue Response Generation**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 8404–8419.

Wei Liu and Michael Strube. 2025c. **Discourse Relation-Enhanced Neural Coherence Modeling**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, pp. 4748–4762.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022c. **P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 61–68.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023f. **G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 2511–2522.

Yang Liu, Ivan Titov, and Mirella Lapata. 2019a. **Single Document Summarization as Tree Induction**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1745–1755.

Ye Liu, Jianguo Zhang, Yao Wan, Congying Xia, Lifang He, and Philip Yu. 2021b. **HET-FORMER: Heterogeneous Transformer with Sparse Attention for Long-Text Extractive Summarization**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 146–154.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024d. **Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators**. In *First Conference on Language Modeling*.

Yixin Liu and Pengfei Liu. 2021c. **SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, pp. 1065–1072.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022d. **BRIO: Bringing Order to Abstractive Summarization**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 2890–2903.

Zechun Liu, Barlas Oguz, Aasish Pappu, Yangyang Shi, and Raghuraman Krishnamoorthi. 2023g. **Binary and Ternary Natural Language Generation**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 65–77.

Zhengyuan Liu and Nancy Chen. 2019b. **Exploiting Discourse-Level Segmentation for Extractive Summarization**. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Hong Kong, China: Association for Computational Linguistics, pp. 116–121.

Zhengyuan Liu and Nancy Chen. 2021d. **Controllable Neural Dialogue Summarization with Personal Named Entity Planning**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 92–106.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2020. **Multilingual Neural RST Discourse Parsing**. In *Proceedings of the 28th International Conference on Computational Linguis-*

tics. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 6730–6738.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021e. **DMRST: A Joint Framework for Document-Level Multilingual RST Discourse Segmentation and Parsing**. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*. Punta Cana, Dominican Republic and Online: Association for Computational Linguistics, pp. 154–164.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. **LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models**. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. St. Julian's, Malta: Association for Computational Linguistics, pp. 139–151.

Ilya Loshchilov and Frank Hutter. 2019. **Decoupled Weight Decay Regularization**. In *International Conference on Learning Representations*.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. **Discourse indicators for content selection in summarization**. In *Proceedings of the SIGDIAL 2010 Conference*. Tokyo, Japan: Association for Computational Linguistics, pp. 147–156.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. **MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts**. In *The Twelfth International Conference on Learning Representations*.

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)*. Vienna, Austria: Association for Computational Linguistics, pp. 10948–10963.

Kun Luo, Zheng Liu, Shitao Xiao, Tong Zhou, Yubo Chen, Jun Zhao, and Kang Liu. 2024. **Landmark Embedding: A Chunking-Free Embedding Method For Retrieval Augmented Long-Context Large Language Models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 3268–3281.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. **Readability Controllable Biomedical Document Summarization**. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 4667–4680.

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)*. Vienna, Austria: Association for Computational Linguistics, pp. 8679–8696.

Xueguang Ma, Kai Sun, Ronak Pradeep, and Jimmy Lin. 2021. **A Replication Study of Dense Passage Retriever**. In *CoRR* abs/2104.05740.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. **Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 12585–12602.

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)*. St. Julian's, Malta: Association for Computational Linguistics, pp. 2803–2815.

Louis Mahon and Mirella Lapata. 2024a. **A Modular Approach for Multimodal Summarization of TV Shows**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 8272–8291.

Louis Mahon and Mirella Lapata. 2024b. **ScreenWriter: Automatic Screenplay Generation and Movie Summarisation**. In *arXiv preprint arXiv:2410.19809*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. **When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories**. In *Proceedings of the 61st Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, pp. 9802–9822.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. Tech. rep. 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*. In *Text-interdisciplinary Journal for the Study of Discourse* 8.3, pp. 243–281.

Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Scott Yih, and Madian Khabza. 2022a. **UniPELT: A Unified Framework for Parameter-Efficient Language Model Tuning**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 6253–6264.

Yuning Mao, Ming Zhong, and Jiawei Han. 2022b. **CiteSum: Citation Text-guided Scientific Extreme Summarization and Domain Adaptation with Limited Supervision**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 10922–10935.

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

Daniel Marcu. 1998. Improving summarization through rhetorical parsing tuning. In *Sixth Workshop on Very Large Corpora*.

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

Daniel Marcu. 1999b. *Discourse trees are good indicators of importance in text*. In *Advances in automatic text summarization* 293, pp. 123–136.

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT press.

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*. Vienna, Austria: Association for Computational Linguistics, pp. 16682–16699.

Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, Chenlin Zhou, Jiayi Mao, Tianze Xia, Jiafeng Guo, and Shenghua Liu. 2025. **A Survey of Context Engineering for Large Language Models**. In *CoRR* abs/2507.13334.

Yisong Miao and Min-Yen Kan. 2025. **Discursive Circuits: How Do Language Models Understand Discourse Relations?** In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, pp. 32570–32589.

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)*. Vienna, Austria: Association for Computational Linguistics, pp. 19019–19035.

Amit Moryossef, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling. 2023. **An Open-Source Gloss-Based Baseline for Spoken to Signed Language Translation**. In *Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages*. Tampere, Finland: European Association for Machine Translation, pp. 22–33.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. **Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1797–1807.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. **Conditional Generation with a Question-Answering Blueprint**. In *Transactions of the Association for Computational Linguistics* 11, pp. 974–996.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. *Planning with Learned Entity Prompts for Abstractive Summarization*. In *Transactions of the Association for Computational Linguistics* 9, pp. 1475–1492.

Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2021. *RST Parsing from Scratch*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 1613–1625.

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*. Vienna, Austria: Association for Computational Linguistics, pp. 13072–13089.

Annishka Oksa, Slava Kalyuga, and Paul Chandler. 2010. *Expertise reversal effect in using explanatory notes for readers of Shakespearean text*. In *Instructional Science* 38, pp. 217–236.

OpenAI. 2023. *GPT-4 Technical Report*. In *CoRR* abs/2303.08774.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022. *The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 5098–5111.

Keighley Overbay, Jaewoo Ahn, Fatemeh Pesaran zadeh, Joonsuk Park, and Gunhee Kim. 2023. *mRedditSum: A Multimodal Abstractive Summarization Dataset of Reddit Threads with Images*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 4117–4132.

Artidoro Pagnoni, Alex Fabbri, Wojciech Kryscinski, and Chien-Sheng Wu. 2023. *Socratic Pretraining: Question-Driven Pretraining for Controllable Summarization*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 12737–12755.

Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. **Multi-modal Abstractive Summarization for How2 Videos**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 6587–6596.

Pinelopi Papalampidi and Mirella Lapata. 2023. **Hierarchical3D Adapters for Long Video-to-text Summarization**. In *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 1297–1320.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a Method for Automatic Evaluation of Machine Translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318.

Vaidehi Patil, Leonardo Ribeiro, Mengwen Liu, Mohit Bansal, and Markus Dreyer. 2024. **REFINESUMM: Self-Refining MLLM for Generating a Multimodal Summarization Dataset**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 13773–13786.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. **YaRN: Efficient Context Window Extension of Large Language Models**. In *The Twelfth International Conference on Learning Representations*.

Jason Phang, Yao Zhao, and Peter Liu. 2023. **Investigating Efficiently Extending Transformers for Long Input Summarization**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 3946–3961.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. **On Extractive and Abstractive Neural Document Summarization with Transformer Language Models**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 9308–9319.

Edoardo Maria Ponti, Alessandro Sordani, Yoshua Bengio, and Siva Reddy. 2023. **Combining Parameter-efficient Modules for Task-level Generalisation**. In *Proceedings of the*

17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 687–702.

Matt Post. 2018. **A Call for Clarity in Reporting BLEU Scores**. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, pp. 186–191.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. **SPIQA: A Dataset for Multimodal Question Answering on Scientific Papers**. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Adithya Pratapa, Kevin Small, and Markus Dreyer. 2023. **Background Summarization of Event Timelines**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 8111–8136.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. **Data-to-Text Generation with Content Selection and Planning**. In *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01, pp. 6908–6915.

Peng Qian, Tahira Naseem, Roger Levy, and Ramón Fernandez Astudillo. 2021. **Structural Guidance for Transformer Language Models**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 3735–3745.

Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, Bo Li, and Lijuan Wang. 2024. **MMSum: A Dataset for Multimodal Summarization and Thumbnail Generation of Videos**. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, pp. 21909–21921.

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*. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 2155–2177.

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). Vienna, Austria: Association for Computational Linguistics, pp. 6982–6995.

Aishwarya Ramakrishnan and Chun-Kit Ngan. 2022. **A Hybrid Video-to-Text Summarization Framework and Algorithm on Cascading Advanced Extractive- and Abstractive-based Approaches for Supporting Viewers' Video Navigation and Understanding**. In *2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pp. 36–39.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. **On Context Utilization in Summarization with Large Language Models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 2764–2781.

Craige Roberts. 2012. *Information structure: Towards an integrated formal theory of pragmatics*. In *Semantics and pragmatics 5*, pp. 6–1.

Stephen Robertson and Hugo Zaragoza. 2009. *The Probabilistic Relevance Framework: BM25 and Beyond*. In *Found. Trends Inf. Retr.* 3.4, pp. 333–389.

Kashob Kumar Roy, Pritom Saha Akash, Kevin Chen-Chuan Chang, and Lucian Popa. 2024. **CONTRGen: Context-driven Tree-structured Retrieval for Open-domain Long-form Text Generation**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, pp. 13773–13784.

Qian Ruan, Malte Ostendorff, and Georg Rehm. 2022. **HiStruct+: Improving Extractive Text Summarization with Hierarchical Structure Information**. In *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 1292–1308.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. **AdapterDrop: On the Efficiency of Adapters in Transformers**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 7930–7946.

Devendra Sachan, Yuhao Zhang, Peng Qi, and William L. Hamilton. 2021. **Do Syntax Trees Help Pre-trained Transformers Extract Information?** In *Proceedings of the 16th*

Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, pp. 2647–2661.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset for Multimodal Language Understanding. In *NeurIPS*.

Diego Sanmartin. 2024. *KG-RAG: Bridging the Gap Between Knowledge and Creativity*. In *CoRR* abs/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*.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. *DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3281–3290.

Sebastian Schuster and Tal Linzen. 2022. *When a sentence does not introduce a discourse entity, Transformer-based models still sometimes refer to it*. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 969–982.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. *MLSUM: The Multilingual Summarization Corpus*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8051–8067.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. *QuestEval: Summarization Asks for Fact-based Evaluation*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 6594–6604.

Alessandro Scirè, Simone Conia, Simone Ciciliano, and Roberto Navigli. 2023. *Echoes from Alexandria: A Large Resource for Multilingual Book Summarization*. In *Findings*

of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics, pp. 853–867.

Encarnación Segarra Soriano, Vicent Ahuir, Lluís-F. Hurtado, and José González. 2022. **DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 5931–5943.

Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. 2021. **Multimodal Video Summarization via Time-Aware Transformers**. In *Proceedings of the 29th ACM International Conference on Multimedia*. MM '21. Virtual Event, China: Association for Computing Machinery, pp. 1756–1765.

Noam Shazeer and Mitchell Stern. 2018. **Adafactor: Adaptive Learning Rates with Sub-linear Memory Cost**. In *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 4596–4604.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. **Multi-LexSum: Real-world Summaries of Civil Rights Lawsuits at Multiple Granularities**. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Wei Shi and Vera Demberg. 2019. **Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 5790–5796.

Omer Shubi, Cfir Avraham Hadar, and Yevgeni Berzak. 2025. **Decoding Reading Goals from Eye Movements**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, pp. 5616–5637.

Woomin Song, Seunghyuk Oh, Sangwoo Mo, Jaehyung Kim, Sukmin Yun, Jung-Woo Ha, and Jinwoo Shin. 2024a. **Hierarchical Context Merging: Better Long Context Un-**

derstanding for Pre-trained LLMs. In *The Twelfth International Conference on Learning Representations*.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024b. **VeriScore: Evaluating the factuality of verifiable claims in long-form text generation**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, pp. 9447–9474.

Neha Srikanth and Junyi Jessy Li. 2021. **Elaborative Simplification: Content Addition and Explanation Generation in Text Simplification**. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 5123–5137.

Aseem Srivastava, Smriti Joshi, Tanmoy Chakraborty, and Md Shad Akhtar. 2024. **Knowledge Planning in Large Language Models for Domain-Aligned Counseling Summarization**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 17775–17789.

Manfred Stede, Maite Taboada, and Debopam Das. 2017. *Annotation guidelines for rhetorical structure*. In *Manuscript*. University of Potsdam and Simon Fraser University.

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*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 8273–8288.

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*. SIGIR '25. Padua, Italy: Association for Computing Machinery, pp. 1240–1250.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. **Document-Level Text Simplification: Dataset, Criteria and Baseline**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 7997–8013.

Ashima Suvarna, Xiao Liu, Tanmay Parekh, Kai-Wei Chang, and Nanyun Peng. 2024. **QUDSELECT: Selective Decoding for Questions Under Discussion Parsing**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 1288–1299.

Shahbaz Syed, Dominik Schwabe, Khalid Al-Khatib, and Martin Potthast. 2023. **Indicative Summarization of Long Discussions**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 2752–2788.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Ponzetto. 2024. **ACLSum: A New Dataset for Aspect-based Summarization of Scientific Publications**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico: Association for Computational Linguistics, pp. 6660–6675.

Chen Tang, Shun Wang, Tomas Goldsack, and Chenghua Lin. 2023a. **Improving Biomedical Abstractive Summarisation with Knowledge Aggregation from Citation Papers**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 606–618.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023b. **Understanding Factual Errors in Summarization: Errors, Summarizers, Datasets, Error Detectors**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 11626–11644.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. **Long Range Arena : A Benchmark for Efficient Transformers**. In *International Conference on Learning Representations*.

Gemini Team. 2023. **Gemini: A Family of Highly Capable Multimodal Models**. In *CoRR* abs/2312.11805.

Llama Team. 2024. **The Llama 3 Herd of Models**. In *CoRR* abs/2407.21783.

- Qwen Team. 2025. *Qwen3 Technical Report*. In *CoRR* abs/2505.09388.
- Katrin Tomanek, Vicky Zayats, Dirk Padfield, Kara Vaillancourt, and Fadi Biadsy. 2021. *Residual Adapters for Parameter-Efficient ASR Adaptation to Atypical and Accented Speech*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 6751–6760.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. *LLaMA: Open and Efficient Foundation Language Models*. In *CoRR* abs/2302.13971.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. *Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 10014–10037.
- Ahmet Üstün and Asa Cooper Stickland. 2022. *When does Parameter-Efficient Transfer Learning Work for Machine Translation?* In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 7919–7933.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention is All you Need*. In *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. *CIDEr: Consensus-based image description evaluation*. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, pp. 4566–4575.
- Jesse Vig and Yonatan Belinkov. 2019. *Analyzing the Structure of Attention in a Transformer Language Model*. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 63–76.

Yixin Wan, Kuan-Hao Huang, and Kai-Wei Chang. 2023. **PIP: Parse-Instructed Prefix for Syntactically Controlled Paraphrase Generation**. In *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 10372–10380.

Bin Wang, Zhengyuan Liu, and Nancy Chen. 2023a. **Instructive Dialogue Summarization with Query Aggregations**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 7630–7653.

Haining Wang, Jason Clark, Hannah McKelvey, Leila Sterman, Zheng Gao, Zuoyu Tian, Sandra Kübler, and Xiaozhong Liu. 2025a. **Science out of its Ivory Tower: improving accessibility with reinforcement learning**. In *Scientometrics* 130.8, pp. 4519–4543.

Jiaan Wang, Fandong Meng, Zengkui Sun, Yunlong Liang, Yuxuan Cao, Jiarong Xu, Haoxiang Shi, and Jie Zhou. 2025b. **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)*. Vienna, Austria: Association for Computational Linguistics, pp. 11328–11344.

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*. Miami, Florida, USA: Association for Computational Linguistics, pp. 5627–5646.

Shu Wang, Yixiang Fang, Yingli Zhou, Xilin Liu, and Yuchi Ma. 2025c. **ArchRAG: Attributed Community-based Hierarchical Retrieval-Augmented Generation**. In *CoRR* abs/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*. Miami, Florida, USA: Association for Computational Linguistics, pp. 17716–17736.

Ye Wang, Xiaojun Wan, and Zhiping Cai. 2022. **Guiding Abstractive Dialogue Summarization with Content Planning**. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 3408–3413.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. **Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 8640–8665.

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. 2025d. **Document Segmentation Matters for Retrieval-Augmented Generation**. In *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria: Association for Computational Linguistics, pp. 8063–8075.

Miriam Wanner, Benjamin Van Durme, and Mark Dredze. 2025. **DnDScore: Decontextualization and Decomposition for Factuality Verification in Long-Form Text Generation**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, pp. 23609–23626.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. **Finetuned Language Models are Zero-Shot Learners**. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Chenxi Whitehouse, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata. 2024. **Low-Rank Adaptation for Multilingual Summarization: An Empirical Study**. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City, Mexico: Association for Computational Linguistics, pp. 1202–1228.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. **NoisyTune: A Little Noise Can Help You Finetune Pretrained Language Models Better**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 680–685.

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)*. Vienna, Austria: Association for Computational Linguistics, pp. 28443–28467.

Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. 2024a. **A Comprehensive Study of Multimodal Large Language Models for Image Quality Assessment**. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXIV*. Vol. 15132. Lecture Notes in Computer Science. Springer, pp. 143–160.

Yating Wu, Ritika Mangla, Alexandros G. Dimakis, Greg Durrett, and Junyi Jessy Li. 2024b. **Which questions should I answer? Saliency Prediction of Inquisitive Questions**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 19969–19987.

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023a. **QUDeval: The Evaluation of Questions Under Discussion Discourse Parsing**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 5344–5363.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023b. **Elaborative Simplification as Implicit Questions Under Discussion**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 5525–5537.

Adrian de Wynter, Xun Wang, Alex Sokolov, Qilong Gu, and Si-Qing Chen. 2023. *An evaluation on large language model outputs: Discourse and memorization*. In *Nat. Lang. Process. J.* 4, p. 100024.

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*.

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2020. **Do We Really Need That Many Parameters In Transformer For Extractive Summarization? Discourse Can Help !** In *Proceedings of the First Workshop on Computational Approaches to Discourse*. Online: Association for Computational Linguistics, pp. 124–134.

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2021. **Predicting Discourse Trees from Transformer-based Neural Summarizers**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 4139–4152.

Qianqian Xie, Jimin Huang, Tulika Saha, and Sophia Ananiadou. 2022. **GRETEL: Graph Contrastive Topic Enhanced Language Model for Long Document Extractive Summarization**. In *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 6259–6269.

Jiacheng Xu and Greg Durrett. 2019. **Neural Extractive Text Summarization with Syntactic Compression**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3292–3303.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. **Discourse-Aware Neural Extractive Text Summarization**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5021–5031.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. **Optimizing Statistical Machine Translation for Text Simplification**. In *Transactions of the Association for Computational Linguistics* 4, pp. 401–415.

Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, XIAOPENG ZHANG, and Qi Tian. 2024. **QA-LoRA: Quantization-Aware Low-Rank Adaptation of Large Language Models**. In *The Twelfth International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. **HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2369–2380.

Zhuoyi Yang, Ming Ding, Yanhui Guo, Qingsong Lv, and Jie Tang. 2022. **Parameter-Efficient Tuning Makes a Good Classification Head**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 7576–7586.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2025. **mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models**. In *The Thirteenth International Conference on Learning Representations*.

Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. **RST Discourse Parsing with Second-Stage EDU-Level Pre-training**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 4269–4280.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. 2024. **MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI**. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, pp. 9556–9567.

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*.

Frances Yung, Kaveri Anuranjana, Merel Scholman, and Vera Demberg. 2022. **Label distributions help implicit discourse relation classification**. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*. Gyeongju, Republic of Korea and Online: International Conference on Computational Linguistics, pp. 48–53.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. **Big Bird: Transformers for Longer Sequences**. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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**. In *Computational Linguistics* 51.1, pp. 23–72.

Guangtao Zeng, Peiyuan Zhang, and Wei Lu. 2023. **One Network, Many Masks: Towards More Parameter-Efficient Transfer Learning**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 7564–7580.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. **Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Singapore: Association for Computational Linguistics, pp. 543–553.

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)*. Vienna, Austria: Association for Computational Linguistics, pp. 7128–7141.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. **PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization**. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 11328–11339.

Litian Zhang, Xiaoming Zhang, Linfeng Han, Zelong Yu, Yun Liu, and Zhoujun Li. 2024a. **Multi-task Hierarchical Heterogeneous Fusion Framework for multimodal summarization**. In *Inf. Process. Manage.* 61.4.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023b. **Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning**. In *The Eleventh International Conference on Learning Representations*.

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)*. Vienna, Austria: Association for Computational Linguistics, pp. 4184–4202.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. **BERTScore: Evaluating Text Generation with BERT**. In *International Conference on Learning Representations*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025c. **Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models**. In *CoRR* abs/2506.05176.

Yu Zhang, Xiushi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024b. **A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 8783–8817.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. **Summ^N: A Multi-Stage Summarization Framework for Long Input Dialogues and Documents**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 1592–1604.

Zhen-Ru Zhang, Chuanqi Tan, Haiyang Xu, Chengyu Wang, Jun Huang, and Songfang Huang. 2023c. **Towards Adaptive Prefix Tuning for Parameter-Efficient Language Model Fine-tuning**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 1239–1248.

Zhihao Zhang, Tomas Goldsack, Carolina Scarton, and Chenghua Lin. 2024c. **ATLAS: Improving Lay Summarisation with Attribute-based Control**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 337–345.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020c. **SG-Net: Syntax-Guided Machine Reading Comprehension**. In *The Thirty-Fourth AAI Conference on Artificial Intelligence, AAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, pp. 9636–9643.

Haodong Zhao, Ruifang He, Mengnan Xiao, and Jing Xu. 2023. **Infusing Hierarchical Guidance into Prompt Tuning: A Parameter-Efficient Framework for Multi-level Implicit Discourse Relation Recognition**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 6477–6492.

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)*. Vienna, Austria: Association for Computational Linguistics, pp. 5172–5189.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. **Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena**. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. **Discourse Level Factors for Sentence Deletion in Text Simplification**. In *The Thirty-Fourth AAI Conference on Artificial Intelligence, AAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, pp. 9709–9716.

Yang Zhong and Diane Litman. 2025a. **A Tale of Evaluating Factual Consistency: Case Study on Long Document Summarization Evaluation**. In *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria: Association for Computational Linguistics, pp. 12511–12532.

Yang Zhong and Diane Litman. 2025b. **Discourse-Driven Evaluation: Unveiling Factual Inconsistency in Long Document Summarization**. 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)*. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 2050–2073.

Jiawei Zhou, Hang Ding, and Haiyun Jiang. 2026. **ARK: Answer-Centric Retriever Tuning via KG-augmented Curriculum Learning**.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. **Towards Automatic Learning of Procedures From Web Instructional Videos**. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, pp. 7590–7598.

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)*. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 8912–8924.

Yunqi Zhu, Xuebing Yang, Yuanyuan Wu, and Wensheng Zhang. 2023. **Parameter-Efficient Fine-Tuning with Layer Pruning on Free-Text Sequence-to-Sequence Modeling**. In *CoRR abs/2305.08285*.

