RST-Enhanced Argument Mining

Prajakta Kini
University of Colorado Boulder
prajakta.kini@colorado.edu

Marshall Washington
University of Colorado Boulder
marshall.washington@colorado.edu

Abstract

This project investigates whether explicit rhetorical structure can improve argument mining. While many argument mining systems rely on surface text, and contextual semantics, argumentation is heavily influenced by discourse relations. To capture these relations, we draw on Rhetorical Discourse Structure (RST). To test whether RST Provides useful structure beyond raw text, we evaluate three RST-enhanced approaches on two datasets. The methods include RST soft prompting for BERT-based classifiers, a relational graph neural network that models EDUs and discourse relations as a graph, and zero-shot LLM inference with RST-annotated inputs. Across both datasets, the most significant results appear in relation classification rather than component classification. Overall, the findings suggest that discourse structure is most valuable when the task relies on relational reasoning, and less helpful when lexical cues already provide strong signals for the model.

1 Introduction

Argument mining is the identification and extraction of the structure of inference and reasoning, as expressed in argumentation. Understanding argument structure makes it possible to determine what positions people are adopting and why. Argument analysis aims to transform unstructured text into structured argument data, providing an understanding of indivdual points, and the relationships between them (Lawrence and Reed, 2019).

Prominent approaches to argument mining include opinion mining, controversy detection, citation mining, and argumentative zoning. Opinion mining, also referred to as sentiment analysis, is the study of opinions, sentiments, and emotions in text (Liu, 2010). An extension of opinion mining, controversy detection aims to execute the automatic detection of controversial topics and conflicting viewpoints. Boltužić and Šnajder (2015) provide the clearest link between argumentation and controversy, where argumentative statements are clustered based on their textual similarity. Citation mining involves the labeling of citations within scientific writing and their rhetorical roles. There are a wide variety of manual schemes for classifying citation motivation and function that can be used for downstream automation (Teufel et al., 2006; Piao et al., 2007; Athar, 2011). Argumentative Zoning (AZ) classifies sentences by their rhetorical and argumentative role within a scientific paper (Teufel et al., 2006).

Argument mining has become an increasingly important task in natural language processing, supporting applications such as automated essay scoring and fact-checking. Despite strong performance of transformer-based models like BERT and RoBERTa, most existing approaches treat text as a flat sequence and rely primarily on lexical and contextual signals (Das et al., 2025). This stands in contrast to human argumentation, which is inherently structured and shaped by discourse-level relationships between clauses and sentences.

Rhetorical Structure Theory (RST) provides a formal framework for capturing these relationships through hierarchical structure, nuclearity, and rhetorical relations (Mann and Thompson, 1987), yet its integration into modern neural argument mining systems remains limited and inconsistent. Prior work that incorporates discourse features often focuses on narrow settings or single modeling paradigms, which makes it unclear whether RST offers generalizable benefits across tasks, domains, and model architectures. There is also little systematic comparison between different strategies for injecting discourse structure, such as input-level prompting, graph-based modeling, and large language model inference, and limited understanding of when and why such structure improves performance. This gap motivates a comprehensive evaluation of RST-enhanced methods to determine

whether explicit discourse information meaningfully improves argument mining beyond what is captured by raw text alone.

2 Related Work

Research on argument mining has focused on identifying claims, premises, and the relations that structure reasoning in text. Early computational approaches often modeled argumentation as a structured prediction problem, including minimum spanning tree (MST)-style parsing over argument graphs (Peldszus and Stede, 2015). Subsequent work explored incorporating discourse-informed features and transferring representations from discourse parsing into argument mining tasks. For example, Hewett et al. (2019) examine automatically derived discourse features for predicting argumentative structure, while Accuosto and Saggion (2019) leverage discourse representations to improve argument mining in scientific text. More recent work further suggests that rhetorical structure can provide a useful inductive bias for modeling argumentation across varied discourse realizations (Chistova, 2023).

Beyond argument mining, a growing body of work integrates Rhetorical Structure Theory into neural architectures that explicitly model discourse structure. Recursive neural models have been used to encode RST trees for downstream tasks such as text classification (Ji and Smith, 2017), and tree-structured models have demonstrated that even automatically generated RST parses can improve coherence modeling (Guz and Carenini, 2020). More recent approaches incorporate rhetorical relations directly into transformer-based architectures. For instance, RST-aware transformer variants encode discourse relations and structural uncertainty to improve summarization and representation learning (Liu et al., 2023), while parameter-efficient methods such as RST-LoRA extend this idea to lightweight adaptation of pretrained models (Liu and Demberg, 2024). In parallel, graph-based formulations have become increasingly prominent, representing discourse as a network of elementary discourse units connected by rhetorical relations. These approaches apply graph neural networks such as GCNs and GATs to capture relational dependencies within discourse structure (Lambropoulos and Ishihara, 2024), aligning closely with broader trends in structured representation learning.

Recent work also explores expressing rhetorical structure in formats compatible with large language models. Prompt-based approaches have been used to encode RST parsing strategies as structured inputs, enabling models such as LLaMA to perform discourse parsing through natural language prompts (Maekawa et al., 2024). Similarly, LLM-based discourse parsers demonstrate that large pretrained models can learn to recover rhetorical structure through fine-tuning or instruction-based inference (Thompson et al., 2024). These approaches suggest that discourse structure can be injected either explicitly through input representations or implicitly through model adaptation. Taken together, this line of work motivates a systematic comparison of input-level prompting, graph-based modeling, and LLM-based inference as alternative strategies for incorporating rhetorical structure into downstream tasks.

3 Methodology

We study discourse-enhanced argument mining across two benchmark settings: (1) persuasive essays (AAEC (Stab and Gurevych, 2017)), and (2) biomedical randomized controlled trial abstracts (AbstRCT (Mayer et al., 2020)). Both datasets are evaluated on two tasks:

Task 1 - Argument Component Classification Given an input sentence s, predict one of: *MajorClaim*, *Claim*, *Premise*, or *O* (non-argumentative).

Because both AAEC and AbstRCT provide gold annotations as character-level argument spans rather than sentence labels, we convert span annotations into sentence-level supervision through span-to-sentence projection. Documents are sentence-tokenized, and each sentence is assigned the label of the argument component with maximum character overlap. Sentences with no overlapping annotated span are labeled *O*.

Formally, for sentence s_k:

$$y_k = \arg\max_{c \in C} \text{OverlapChars}(s_k, c) \qquad (1)$$

where C is the set of annotated component spans. If no overlap exists, $y_k = O$.

This strategy standardizes Task 1 across datasets while preserving compatibility with original span-level annotations.

Task 2 - Argumentative Relation Classification: Given an ordered pair of annotated argument spans (a_i, a_j), predict their argumentative relation.

AAEC uses *supports*, *attacks*, or *none*, while AbstRCT uses *Support*, *Attack*, *Partial-Attack*, or *none* (Attack and Partial-Attack merged into a single class).

For Task 2, positive pairs are derived directly from gold relation annotations, while negative pairs are sampled at a 1:1 ratio from unannotated but structurally plausible component pairs. Class imbalance is handled via inverse-frequency class weights in the cross-entropy loss for all fine-tuned models.

3.1 RST parsing

All approaches use the DMRST parser (Liu et al., 2021), a hierarchical neural RST parser based on XLM-RoBERTa. RST parses are computed once per document and cached to disk. The parser produces Elementary Discourse Units (EDUs) with character offsets and a constituency tree encoding nuclearity (Nucleus/Satellite) and relation type for each edge.

Five relation types are observed in practice on both datasets: *Joint*, *Elaboration*, *Background*, *Explanation*, and *Temporal*.

3.2 Approach 1: BERT/RoBERTa Baseline

We fine-tune BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) independently on Task 1 and Task 2.

Task 1. Each sentence is encoded as:

$$h = \text{Encoder}([\texttt{CLS}]\ s\ [\texttt{SEP}]) \tag{2}$$

Classification is performed using the final [CLS] representation:

$$\hat{y} = \text{softmax}(W h_{\texttt{CLS}}) \tag{3}$$

Task 2. Each span pair is encoded using standard sentence-pair input:

$$h = \text{Encoder}([\texttt{CLS}]\ a_i\ [\texttt{SEP}]\ a_j\ [\texttt{SEP}]) \tag{4}$$

A classification head predicts the relation label.

Both tasks use AdamW with learning rate 2×10^{-5}, dropout 0.3, weight decay 0.1, and early stopping with patience 5 on macro-F1.

3.3 Approach 2: RST Soft Prompting

We extend the baseline by restructuring input text with RST discourse markup, injecting rhetorical structure directly into the Transformer input sequence through 24 newly introduced special tokens (10 structural tokens and 14 relation tokens). This enables pretrained language models to internalize discourse organization without requiring architectural modification, preserving the standard BERT/RoBERTa encoder while enriching inputs with explicit rhetorical signals.

Structural tokens (10)

```
<NUCLEUS>, </NUCLEUS>, <SATELLITE>,
</SATELLITE>, <RELATION>, </RELATION>,
<DISCOURSE>, </DISCOURSE>, <NO_RST>,
</NO_RST>
```

RST relation tokens (14)

```
<Attribution>, <Background>, <Cause>,
<Comparison>, <Concession>, <Condition>,
<Contrast>, <Elaboration>, <Enablement>,
<Evidence>, <Explanation>, <Joint>,
<Summary>, <Temporal>
```

Each sentence is restructured at three levels of discourse structure:

(1) Intra-sentence nuclearity: each overlapping EDU is wrapped in nuclearity tags.

(2) Intra-sentence relations: <RELATION> tags are inserted between consecutive EDU pairs with a known RST relation.

(3) Cross-sentence discourse: a <DISCOURSE> tag encodes the RST relation between the last EDU of this sentence and the first EDU of the next sentence.

For Task 2, only nuclearity tags are used; relation and discourse tags are excluded to prevent label leakage, as the inter-span RST relation is the label being predicted.

Sentences with no overlapping EDUs receive a <NO_RST> tag. Special token embeddings are randomly initialized and fine-tuned jointly with BERT. Hyperparameters follow Approach 1.

3.4 Approach 3: R-GCN over RST Discourse Graph

We model document-level discourse structure as a relational graph, where EDUs are nodes and RST relations are typed directed edges. Each graph contains 11 edge types: 5 forward relations, 5 reverse relations, and 1 self-loop.

Node initialization. Each EDU is encoded using frozen BERT.

R-GCN layers. Two stacked R-GCN layers propagate discourse information:

$$\mathbf{h}_i^{(l+1)} = \text{ReLU}\Bigg(\mathbf{W}_{\text{self}}\mathbf{h}_i^{(l)} + \sum_{r=1}^{R} \frac{\alpha_r^{\text{dir}}}{|\mathcal{N}_r(i)|} \sum_{j \in \mathcal{N}_r(i)} \mathbf{W}_r \mathbf{h}_j^{(l)}\Bigg) \quad (5)$$

where α_r^{dir} are learnable nuclearity scalars for satellite→nucleus, nucleus→satellite, and multinuclear edges.

Here, $\mathbf{h}_i^{(l)}$ is the representation of EDU node i at layer l. Each R-GCN layer updates node i by combining its own transformed representation, $\mathbf{W}_{\text{self}}\mathbf{h}_i^{(l)}$, with normalized messages from discourse neighbors $\mathcal{N}_r(i)$ across all relation types r. Relation-specific matrices $\mathbf{W}_r$ model different discourse relations, while α_r^{dir} are learnable scalars that weight edges based on nuclearity direction. Stacking two layers allows each EDU to capture both local and broader RST discourse context.

Task 1. Sentence representations are obtained by mean-pooling all EDU embeddings overlapping the sentence:

$$\mathbf{h}_s = \frac{1}{|\mathcal{E}_s|} \sum_{i \in \mathcal{E}_s} \mathbf{h}_i^{(2)} \quad (6)$$

Task 2. For each span pair (A, B), pooled span embeddings are combined as:

$$\mathbf{h}_{\text{pair}} = [\mathbf{h}_A;\, \mathbf{h}_B;\, \mathbf{h}_A - \mathbf{h}_B;\, \mathbf{h}_A \odot \mathbf{h}_B] \quad (7)$$

The direct RST edge between spans is excluded to prevent label leakage.

Training. Both tasks are jointly optimized:

$$\mathcal{L} = 0.5\mathcal{L}_{\text{T1}} + 0.5\mathcal{L}_{\text{T2}} \quad (8)$$

This framework integrates discourse structure explicitly while preserving shared representations across argument component and relation prediction.

3.5 Approach 4: GPT-4o Inference Baseline

We evaluate GPT-4o under RST-augmented prompting (same as approach 2) in the zero-shot setting.

RST Mode. Inputs use the same RST markup format as the soft prompt baseline.

This setup allows direct comparison between: (1) pretrained LLM reasoning, (2) discourse-aware prompting, and (3) supervised fine-tuning.

Gold labels are derived identically to the fine-tuned baselines, ensuring direct comparability.

4 Results & Discussion

4.1 RST Benefits Are Strongly Task- and Dataset-Dependent

Our results show that the effectiveness of RST discourse integration depends heavily on both task formulation and dataset domain (Tables 1 - 4; 2-4 can be found in the appendix)

For AAEC Task 2, RST soft prompting produces the largest gains in the entire study, improving macro-F1 by +14.10% for BERT and +13.44% for RoBERTa (Table 2). These substantial improvements indicate that rhetorical structure is highly informative for argumentative relation prediction in persuasive essays, where discourse organization is often implicit and structurally diverse.

In contrast, on AbstRCT Task 2, RST prompting provides negligible or negative effects, slightly reducing performance relative to already strong baselines (Table 3). This suggests that biomedical abstracts contain sufficiently explicit lexical and structural cues such that pretrained transformers already capture most relation information without requiring additional discourse markup.

For Task 1, RST prompting yields only marginal gains overall. AbstRCT BERT improves modestly (+1.11 macro-F1) (Table 1), while AAEC performance is nearly unchanged or slightly degraded (Table 4). These findings imply that argument component classification is driven primarily by lexical semantics, limiting the added value of rhetorical structure.

4.2 RST Particularly Improves Rare Relation Classes in AAEC

RST prompting provides especially large improvements for minority relation classes, most notably the *attacks* class in AAEC (Table 2). BERT improves from 3.64 to 30.56 F1, while RoBERTa rises from 20.59 to 49.32. These gains suggest that nuclearity and rhetorical asymmetry offer valuable structural cues for identifying infrequent argumentative relations that are difficult to infer from lexical content alone.

4.3 R-GCN Underperforms Across Most Settings

Despite its explicit discourse graph formulation, R-GCN generally underperforms both transformer

baselines and RST soft prompting across nearly all settings. Performance declines are particularly severe for Task 1 on both datasets, and Task 2 improvements are limited to only modest gains in AAEC BERT (Tables 4 & 2). This suggests that frozen BERT node embeddings, parser noise, sparse graph connectivity, and loss of fine-grained lexical information collectively hinder graph effectiveness. Under current conditions, lightweight prompt-based RST integration appears substantially more robust than hard graph structuring.

4.4 LLMs Remain Limited for Fine-Grained Argument Mining

GPT-4o performs poorly on Task 1 across both datasets, particularly on AAEC where performance collapses dramatically (Tables 4 and 1). This indicates that zero-shot LLMs struggle with fine-grained argument component distinctions and dataset-specific labeling schemes. On Task 2, however, GPT-4o is comparatively more competitive, slightly surpassing the BERT baseline on AAEC while still remaining well below supervised RST-enhanced models (Table 2). Relation prediction appears better aligned with general reasoning capabilities acquired during large-scale pretraining.

4.5 Dataset Structure Strongly Shapes Discourse Utility

AAEC and AbstRCT differ substantially in discourse organization. AAEC essays are typically less standardized and rhetorically more diverse, making explicit discourse modeling particularly valuable. In contrast, AbstRCT abstracts follow more regular scientific writing patterns and exhibit stronger lexical and structural consistency, allowing raw transformer models to perform strongly even without explicit discourse augmentation. These differences suggest that discourse enhancement methods are most useful when argumentative structure is implicit, diverse, or weakly signaled by surface text.

Overall, our findings indicate that RST discourse information is most valuable for relation classification in structurally complex domains, with soft prompting offering the best performance-to-complexity tradeoff. Graph-based discourse modeling remains promising but currently limited by parser quality and representation constraints, while LLMs still require task-specific supervision for reliable fine-grained argument mining.

Enc.	Meth.	MF1	MC F1	C F1	Pr F1	O F1
BERT	Base	68.95	28.57	69.57	86.34	91.30
	RST	70.06 (+1.11)	25.53 (−3.04)	75.00 (+5.43)	87.82 (+1.48)	91.88 (+0.58)
	R-GCN	64.79 (−4.16)	25.81 (−2.76)	69.12 (−0.45)	80.00 (−6.34)	84.23 (−7.07)
RoBERTa	Base	69.96	26.67	73.81	86.88	92.47
	RST	68.86 (−1.10)	22.22 (−4.45)	74.24 (+0.43)	87.06 (+0.18)	91.93 (−0.54)
	R-GCN	50.92 (−19.04)	20.62 (−6.05)	55.23 (−18.58)	64.92 (−21.96)	62.89 (−29.58)
LLM	GPT-4o	57.49 (−11.46)	08.70 (−19.87)	56.38 (−13.19)	79.79 (−6.55)	85.11 (−6.19)

Table 1: **AbstRCT (Neoplasm) Task 1 : Component Classification** (test set: 1,258 sentences). Macro-F1 and per-class F1 scores are reported as percentage (%). Delta values indicate change relative to the baseline of the corresponding encoder; green denotes improvement and red denotes degradation. LLM (GPT-4o) deltas are computed relative to the BERT baseline. MF1 = Macro-F1; MC = MajorClaim; C = Claim; Pr = Premise; RoB = RoBERTa.

5 Error Analysis

To move beyond aggregate F1 scores and understand the conditions under which RST integration is useful, we conduct a fine-grained error analysis comparing RST soft prompting (RoBERTa) against the RoBERTa baseline across both datasets and both tasks. Each aligned prediction pair is categorized into one of five disagreement categories: RST Fixed (RST correct, baseline wrong), RST Broke (baseline correct, RST wrong), Both Wrong Diff (both incorrect with different predictions), Both Wrong Same (both incorrect with the same prediction), and Both Correct. We additionally conduct per-label benefit analysis and per-document accuracy delta analysis to identify correlations between document-level characteristics and RST's tendency to help or hurt.

5.1 Component Classification (Task 1)

AbstRCT. On AbstRCT Task 1, RST fixes more predictions than it breaks (81 fixed, 65 broke), though the overall disagreement rate is low: 81.6% of all 1,725 aligned pairs are classified correctly by both models, reflecting the already strong baseline performance in this domain (Table 5). The net benefit is therefore modest in absolute terms, consistent with the small macro-F1 gain of +1.10 reported in the main results.

The per-label breakdown (Table 6) reveals a striking asymmetry in where RST helps and hurts. RST strongly benefits *Claim* classification (42 fixed, 3 broke, net +39) while degrading *Premise*

classification (10 fixed, 31 broke, net −21). *MajorClaim* also sees net degradation (−6), though from a very small total (27 instances). The *O* class (non-argumentative) shows a modest net gain (+4).

This asymmetry is interpretable in terms of RST nuclearity. In biomedical abstracts, claims tend to appear as discourse nuclei, as they are the central, rhetorically prominent units that satellite EDUs elaborate or justify. RST nuclearity markup thus provides a useful signal for promoting nucleus-tagged sentences to Claim. However, this same tendency causes RST to over-promote nucleus-tagged premises to Claim. In the Biomedical domain premises are frequently short, declarative, nuclear statements (e.g., *"Compared with MA, there were similar or greater improvements in pain, TRSS, and QOL with EXE."*) that RST cannot distinguish from claims on structural grounds alone. The model appears to over-rely on nuclearity as a proxy for argumentative centrality.

AAEC. On AAEC Task 1, the disagreement profile is more balanced: 94 fixed versus 87 broke, with a higher proportion of persistent errors (Both Wrong Same: 19.8%, compared to 8.2% on AbstRCT), reflecting the overall harder nature of argument component classification in persuasive essays (Table 9).

The per-label breakdown (Table 10) shows the opposite pattern from AbstRCT. RST strongly benefits *Premise* (63 fixed, 32 broke, net +31) and modestly helps *MajorClaim* (+3) and *O* (+13), while substantially hurting *Claim* (6 fixed, 46 broke, net −40). This reversal is consistent with the structural properties of persuasive essays. In AAEC, premises are often multi-EDU constructions with complex internal discourse structure, such as elaborations, explanations, and joint relations, which RST nuclearity helps the model recognize as supporting evidence rather than central assertions. Claims, however, are frequently short single-EDU sentences that appear as nuclei, causing RST to suppress them toward O when their RST context is ambiguous. Qualitative inspection confirms this pattern: broken Claim cases include sentences like *"The most effective way to convince consumers to purchase a product is through advertising,"* where RST fragments the sentence across multiple nucleus spans and the model loses confidence in the Claim label. This suggests that RST's granular EDU-level markup introduces noise for short, syntactically simple claims that are already well-handled by the baseline's contextual embeddings.

The fact that RST helps Claims in one domain but hurts them in the other, and vice versa for Premises points to a domain-specific interaction between discourse structure and argumentative function. Biomedical abstracts follow rigid rhetorical conventions that align nucleus position with claim-level content, while persuasive essays exhibit more varied and less predictable mappings between nuclearity and argumentative role.

5.2 Relation Classification (Task 2)

AbstRCT. RST provides substantially stronger and more consistent benefits on Task 2. On AbstRCT, RST fixes 160 pairs and breaks 72, with a fixed-to-broke ratio of 2.2:1 which is considerably higher than the 1.25:1 ratio on Task 1 (Table 7). This is consistent with the hypothesis that relational reasoning benefits more from explicit discourse structure than per-sentence labeling.

The breakdown per-label (Table 8 shows the largest net gain on *Supports* (+73, with 89 fixed and 16 broken), indicating that RST most consistently helps the model identify supportive relations that the baseline misses. The \textit{Attacks} class sees a net gain of +11 despite its small total (82 pairs), consistent with the finding in Section 4.2 that RST provides useful signal for minority class identification. The \textit{None} class shows a near-zero net gain (+4, with 58 fixed and 54 broke), suggesting RST provides limited disambiguating power for non-relations in biomedical text, likely because topic coherence between adjacent sentences in biomedical abstracts is high regardless of whether an argumentative relation holds, making it difficult for discourse markup alone to reliably distinguish None from Supports.

AAEC. AAEC Task 2 shows the largest disagreement volume in the study: 337 fixed and 201 broke out of 1,618 pairs, with a fixed-to-broke ratio of 1.68:1 (Table 11). The proportion of RST-impacted predictions (33.2% of all pairs) is substantially higher than any other setting, underscoring how consequential RST markup is for relation classification in persuasive essays.

The per-label breakdown (Table 12) shows a massive net gain on *Supports* (+128, with 191 fixed and 63 broke), driven by RST's ability to identify rhetorical elaboration and explanation relations as markers of argumentative support. Many baseline errors on this class involve pairs where the

model predicts *None* for structurally connected sentence pairs. RST largely corrects by making the discourse connection explicit. The *Attacks* class again improves consistently (net +9), confirming RST's value for adversarial relation detection across domains.

The *None* class presents the most interesting pattern: 135 fixed but 136 broke, for a net of -1 (Table 12). This near-perfect cancellation indicates that RST simultaneously helps and hurts *None* detection in roughly equal measure. The core problem appears to be that in persuasive essays, many adjacent sentence pairs are discourse-connected (via joint or elaboration relations) regardless of whether they stand in an argumentative relation. RST markup thus sometimes promotes *None* pairs to *Supports* based on surface discourse connectivity, introducing false positives that cancel out the true corrections. This represents a fundamental limitation: RST discourse relations do not map cleanly onto argumentative relations, and in a domain with dense discourse connectivity, the signal can be as misleading as it is helpful.

5.3 Per-Document Accuracy Delta Analysis

Correlation between baseline accuracy and RST impact. A key question is whether RST's tendency to help or hurt correlates with document-level characteristics. In particular, whether documents where the baseline is already strong are more likely to be hurt by RST, and vice versa.

Examining the per-document delta data for AbstRCT Task 1, there is a clear pattern among the documents most hurt by RST: nearly all had high baseline accuracy before RST was applied. The top hurt documents had baseline accuracies of 0.889, 0.929, 0.833, 0.917, 0.857, 0.875, 1.000, 1.000, 0.944, and 0.700 respectively, with a mean baseline accuracy of 0.892. By contrast, the top improved documents had baseline accuracies of 0.750, 0.750, 0.769, 0.643, 0.600, 0.833, 0.846, 0.846, 0.714, and 0.875, with a mean of 0.763. This difference of approximately 13 percentage points suggests a meaningful negative correlation between baseline accuracy and RST benefit on this task: RST tends to help documents where the baseline already struggles, and hurt documents where it already performs well.

This pattern is consistent with a ceiling effect interpretation. Documents where the baseline achieves near-perfect accuracy have little room for RST to add signal, and any RST-induced perturbation to the input representation is likely to introduce noise rather than useful structure. In contrast, documents where the baseline is confused, likely those with more ambiguous argumentative boundaries or less stereotypical sentence structure, benefit from the additional discourse scaffolding RST provides.

The same correlation holds on AAEC Task 1. The top hurt documents had mean baseline accuracy of approximately 0.810 (ranging from 0.938 to 0.650), while the top improved documents had mean baseline accuracy of approximately 0.598 (ranging from 0.600 to 0.739). Again, RST's benefit is concentrated among lower-performing baseline documents.

AbstRCT Task 2 per-document analysis. On AbstRCT Task 2, the per-document delta pattern is more extreme. The top improved documents include `test_0107`, where the baseline scored 0.000 and RST achieved 1.000, as well as several documents improving from 0.438–0.625 to 0.875–1.000. The documents most hurt by RST had uniformly high baseline scores (0.917, 1.000, 1.000, 0.500, 0.500, 0.917, 1.000, 0.900, 0.833, 0.917), again suggesting the same ceiling effect. The mean baseline accuracy for the top 10 hurt documents is approximately 0.808, compared to approximately 0.476 for the top 10 improved documents. This is a gap of over 33 percentage points, the largest such gap across all four settings.

AAEC Task 2 per-document analysis. AAEC Task 2 shows the largest absolute deltas in the study. The most improved document (`test_0001`) moves from 0.250 to 0.750 (+0.500), and several documents achieve perfect RST accuracy from moderate baselines. The mean baseline accuracy of the top 10 improved documents is approximately 0.461, while the top 10 hurt documents had mean baseline accuracy of approximately 0.718, which is consistent with the help/hurt correlation observed across all settings.

Notably, on AAEC Task 2, even the most hurt documents do not fall below 0.400 RST accuracy, suggesting RST does not catastrophically degrade any individual document's relation classification, but rather redistributes errors from already-weak documents to already-strong ones.

5.4 Summary and Implications

Across all four task-dataset combinations, several consistent patterns emerge from the error analysis.

First, RST's benefit is strongly correlated with baseline difficulty at the document level: documents where the baseline model already performs well are systematically more likely to be hurt by RST, while documents where the baseline struggles are more likely to benefit. This suggests that RST markup functions as a form of additional inductive bias that is helpful when the model lacks sufficient signal but harmful when it introduces perturbation to already well-learned representations. Future work should investigate whether this effect can be predicted in advance. For instance, using baseline confidence scores to selectively apply RST markup only to documents or instances where the model is uncertain.

Second, RST is more reliably beneficial for relation classification than component classification. Fixed-to-broke ratios are consistently higher on Task 2 (2.2:1 on AbstRCT, 1.68:1 on AAEC) than on Task 1 (1.25:1 on AbstRCT, 1.08:1 on AAEC). This aligns with the intuition that discourse structure encodes relational information more directly than per-sentence argumentative roles, which are more strongly determined by lexical content.

Third, the Attacks class benefits consistently across all settings despite its small size, suggesting that RST rhetorical asymmetry provides a reliable signal for adversarial relation detection that is not easily captured by lexical or contextual cues alone. This is one of the clearest and most consistent wins for RST across the study.

Fourth, the None class is the most difficult for RST to improve reliably. In both relation classification tasks, None shows near-zero net benefit, with large numbers of both fixed and broke predictions canceling out. This reflects the fundamental ambiguity of non-relations in discourse-rich text: RST connectivity does not imply argumentative relation, but the model appears prone to inferring one from the other. Better disambiguation between discourse and argumentative structure may require explicit negative training signal or additional features beyond nuclearity and relation type.

6 Conclusion

This paper investigated whether explicit rhetorical structure, as captured by RST, can improve argument mining across two datasets and two tasks. Our results show that the value of discourse structure is highly dependent on both task formulation and domain. RST soft prompting provides the most consistent gains, particularly for relation classification in persuasive essays, where argumentative structure is implicit and rhetorically diverse. In contrast, R-GCN consistently underperforms, suggesting that hard graph-based discourse modeling is currently limited by parser noise and representation constraints. GPT-4o remains competitive on relation classification but struggles with fine-grained component distinctions, indicating that zero-shot LLMs are not yet a reliable substitute for supervised models on these tasks.

Our error analysis reveals that RST's benefit is strongly concentrated among documents and instances where the baseline already struggles, while high-performing baseline documents are more likely to be hurt by RST markup. This ceiling effect suggests that indiscriminate application of discourse structure is not always beneficial, and that future systems may benefit from selectively applying RST based on model confidence. Overall, our findings suggest that discourse structure is most valuable as a complementary signal for relational reasoning, and least helpful when lexical cues already provide strong classification signal.

7 Future Work

Future work should evaluate oracle settings using gold argument components for Task 2 to better isolate relation classification performance, as well as cross-domain transfer to assess generalization across argumentative domains. Promising directions also include discourse-supervised LLM fine-tuning, and hybrid architectures that combine soft prompting with graph reasoning. Finally, deeper R-GCN variants may better capture long-range discourse dependencies than the current two-layer design. Future work should also include more robust error analysis that covers multiple approaches, especially in areas where R-GCN and LLM inference approaches performed poorly.

References

Pablo Accuosto and Horacio Saggion. 2019. Transferring knowledge from discourse to arguments: A case study with scientific abstracts. In *Proceedings of the 6th Workshop on Argument Mining*, pages 41–51, Florence, Italy. Association for Computational Linguistics.

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, pages 81–87, Portland, OR, USA. Association for Computational Linguistics.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, Denver, CO. Association for Computational Linguistics.

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

Nilmadhab Das, Vishal Vaibhav, Yash Sunil Choudhary, V. Vijaya Saradhi, and Ashish Anand. 2025. End-to-end argument mining through autoregressive argumentative structure prediction. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Grigorii Guz and Giuseppe Carenini. 2020. Towards domain-independent text structuring trainable on large discourse treebanks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3141–3152, Online. Association for Computational Linguistics.

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy. Association for Computational Linguistics.

Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005, Vancouver, Canada. Association for Computational Linguistics.

Michael Lambropoulos and Shunichi Ishihara. 2024. Towards an implementation of Rhetorical Structure Theory in discourse coherence modelling. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 1–11, Canberra, Australia. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Bing Liu. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing*.

Dongqi Liu and Vera Demberg. 2024. RST-LoRA: A discourse-aware low-rank adaptation for long document abstractive summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2200–2220, Mexico City, Mexico. Association for Computational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

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

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

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. *Transformer-Based Argument Mining for Healthcare Applications*, pages 2108–2115. SAGE Publications.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Scott Piao, Sophia Ananiadou, Yoshimasa Tsuruoka, Yutaka Sasaki, and John McNaught. 2007. Mining opinion polarity relations of citations.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

Kate Thompson, Akshay Chaturvedi, Julie Hunter, and Nicholas Asher. 2024. Llamipa: An incremental discourse parser. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6418–6430, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

A.1 Task 1 & 2 Results

Enc.	Meth.	MF1	Sup F1	Att F1	None F1
BERT	Base	41.75	57.69	03.64	63.91
	RST	55.85 (+14.10)	70.29 (+12.60)	30.56 (+26.92)	66.71 (+2.80)
	R-GCN	45.67 (+3.92)	66.36 (+8.67)	12.80 (+9.16)	57.84 (−6.07)
RoBERTa	Base	48.95	63.26	20.59	63.00
	RST	62.39 (+13.44)	69.09 (+5.83)	49.32 (+28.73)	68.75 (+5.75)
	R-GCN	43.01 (−5.94)	65.29 (+2.03)	08.47 (−12.12)	55.27 (−7.73)
LLM	GPT-4o	42.44 (+0.69)	67.21 (+9.52)	21.59 (+17.95)	38.52 (−25.39)

Table 2: **AAEC Task 2 : Relation Classification** (test set: 1,618 pairs).

Enc.	Meth.	MF1	Sup F1	Att F1	None F1
BERT	Base	72.28	73.49	69.09	74.27
	RST	68.58 (−3.70)	72.89 (−0.60)	60.55 (−8.54)	72.29 (−1.98)
	R-GCN	56.09 (−16.19)	65.41 (−8.08)	46.90 (−22.19)	55.97 (−18.30)
RoB	Base	71.37	73.71	69.35	71.04
	RST	71.00 (−0.37)	74.06 (+0.35)	66.14 (−3.21)	72.80 (+1.76)
	R-GCN	52.66 (−18.71)	60.91 (−12.80)	48.33 (−21.02)	48.74 (−22.30)
LLM	GPT-4o	52.09 (−20.19)	69.95 (−3.54)	33.96 (−35.13)	52.34 (−21.93)

Table 3: **AbstRCT (Neoplasm) Task 2 : Relation Classification** (test set: 848 pairs; Support n=364, Attack n=60, None n=424). Attack and Partial-Attack labels are merged into a single Attack class.

Enc.	Meth.	MF1	MC F1	C F1	Pr F1	O F1
BERT	Base	68.60	74.68	51.70	76.87	71.14
	RST	68.65 (+0.05)	74.12 (−0.56)	50.97 (−0.73)	77.91 (+1.04)	71.60 (+0.46)
	R-GCN	52.87 (−15.73)	44.27 (−30.41)	38.08 (−13.62)	72.19 (−4.68)	56.95 (−14.19)
RoBERTa	Base	70.73	73.72	56.17	78.67	74.38
	RST	69.66 (−1.07)	74.92 (+1.20)	55.02 (−1.15)	77.03 (−1.64)	71.68 (−2.70)
	R-GCN	48.06 (−22.67)	37.56 (−36.16)	38.72 (−17.45)	65.95 (−12.72)	50.00 (−24.38)
LLM	GPT-4o	14.45 (−54.15)	24.08 (−50.60)	02.11 (−49.59)	00.84 (−76.03)	30.77 (−40.37)

Table 4: **AAEC Task 1 : Component Classification** (test set: 1,381 sentences).

A.2 Error Analysis Results

Category	Count	%
RST Fixed	81	4.7%
RST Broke	65	3.8%
Both Wrong Diff	31	1.8%
Both Wrong Same	141	8.2%
Both Correct	1407	81.6%

Table 5: **AbstRCT Task 1 : Component Classification — Disagreement Analysis** (total aligned pairs: 1,725). RST Fixed = RST correct, Baseline wrong; RST Broke = Baseline correct, RST wrong.

Label	Fixed	Broke	Net	Total
MajorClaim	1	7	−6	27
Claim	42	3	+39	254
Premise	10	31	−21	560
O	28	24	+4	884

Table 6: **AbstRCT Task 1 : Component Classification — Per-Label RST Benefit.** Fixed = cases where RST corrected the baseline; Broke = cases where RST degraded the baseline. Positive Net indicates RST helped more than it hurt for that label.

Category	Count	%
RST Fixed	160	14.1%
RST Broke	72	6.3%
Both Wrong Diff	10	0.9%
Both Wrong Same	112	9.9%
Both Correct	782	68.8%

Table 7: **AbstRCT Task 2 : Relation Classification — Disagreement Analysis** (total aligned pairs: 1,136).

Label	Fixed	Broke	Net	Total
Supports	89	16	+73	485
Attacks	13	2	+11	82
None	58	54	+4	569

Table 8: **AbstRCT Task 2 : Relation Classification — Per-Label RST Benefit.** Positive Net indicates RST helped more than it hurt for that label.

Category	Count	%
RST Fixed	94	6.8%
RST Broke	87	6.3%
Both Wrong Diff	26	1.9%
Both Wrong Same	274	19.8%
Both Correct	900	65.2%

Table 9: **AAEC Task 1 : Component Classification — Disagreement Analysis** (total aligned pairs: 1,381).

Label	Fixed	Broke	Net	Total
MajorClaim	7	4	+3	153
Claim	6	46	−40	274
Premise	63	32	+31	710
O	18	5	+13	244

Table 10: **AAEC Task 1 : Component Classification — Per-Label RST Benefit.** Positive Net indicates RST helped more than it hurt for that label.

Category	Count	%
RST Fixed	337	20.8%
RST Broke	201	12.4%
Both Wrong Diff	21	1.3%
Both Wrong Same	285	17.6%
Both Correct	774	47.8%

Table 11: **AAEC Task 2 : Relation Classification — Disagreement Analysis** (total aligned pairs: 1,618).

Label	Fixed	Broke	Net	Total
Supports	191	63	+128	767
Attacks	11	2	+9	42
None	135	136	−1	809

Table 12: **AAEC Task 2 : Relation Classification — Per-Label RST Benefit.** Positive Net indicates RST helped more than it hurt for that label.

Doc ID	Δ	Base	RST	N
Most Improved				
test_0006	+0.250	0.750	1.000	8
test_0015	+0.250	0.750	1.000	8
test_0027	+0.231	0.769	1.000	13
test_0079	+0.214	0.643	0.857	14
test_0066	+0.200	0.600	0.800	10
test_0109	+0.167	0.833	1.000	12
test_0042	+0.154	0.846	1.000	13
test_0051	+0.154	0.846	1.000	13
test_0008	+0.143	0.714	0.857	7
test_0065	+0.125	0.875	1.000	8
Most Hurt				
test_0080	-0.222	0.889	0.667	9
test_0060	-0.214	0.929	0.714	14
test_0041	-0.167	0.833	0.667	12
test_0030	-0.167	0.917	0.750	12
test_0110	-0.143	0.857	0.714	14
test_0089	-0.125	0.875	0.750	16
test_0062	-0.111	1.000	0.889	9
test_0013	-0.111	1.000	0.889	9
test_0118	-0.111	0.944	0.833	18
test_0114	-0.100	0.700	0.600	10

Table 13: **Per-document accuracy change (Δ) for ABSTRCT Task 1 (component classification). Positive values indicate improvement with RST.**

Doc ID	Δ	Base	RST	N
Most Improved				
test_0107	+1.000	0.000	1.000	1
test_0014	+0.500	0.500	1.000	8
test_0032	+0.500	0.500	1.000	6
test_0133	+0.500	0.500	1.000	10
test_0036	+0.450	0.400	0.850	20
test_0052	+0.438	0.438	0.875	16
test_0011	+0.400	0.600	1.000	10
test_0088	+0.400	0.600	1.000	10
test_0079	+0.375	0.625	1.000	8
test_0081	+0.375	0.500	0.875	8
Most Hurt				
test_0071	-0.333	0.917	0.583	12
test_0132	-0.250	1.000	0.750	4
test_0117	-0.250	1.000	0.750	12
test_0101	-0.250	0.500	0.250	4
test_0047	-0.250	0.500	0.250	4
test_0046	-0.250	0.917	0.667	12
test_0045	-0.250	1.000	0.750	4
test_0082	-0.200	0.900	0.700	10
test_0093	-0.167	0.833	0.667	6
test_0126	-0.167	0.917	0.750	12

Table 14: **Per-document accuracy change (Δ) for ABSTRCT Task 2 (relation classification).**

Doc ID	Δ	Base	RST	N
Most Improved				
test_0008	+0.200	0.600	0.800	15
test_0038	+0.174	0.739	0.913	23
test_0073	+0.154	0.462	0.615	13
test_0002	+0.143	0.714	0.857	21
test_0049	+0.133	0.333	0.467	15
test_0060	+0.130	0.565	0.696	23
test_0047	+0.125	0.688	0.812	16
test_0054	+0.125	0.812	0.938	16
test_0059	+0.105	0.789	0.895	19
test_0016	+0.100	0.500	0.600	10
Most Hurt				
test_0010	-0.250	0.938	0.688	16
test_0051	-0.200	1.000	0.800	15
test_0003	-0.158	0.737	0.579	19
test_0025	-0.154	0.692	0.538	13
test_0032	-0.136	0.773	0.636	22
test_0064	-0.133	0.733	0.600	15
test_0012	-0.133	0.667	0.533	15
test_0036	-0.111	0.722	0.611	18
test_0033	-0.100	0.650	0.550	20
test_0056	-0.091	0.818	0.727	22

Table 15: **Per-document accuracy change (Δ) for AAEC Task 1.**

Doc ID	Δ	Base	RST	N
Most Improved				
test_0001	+0.500	0.250	0.750	12
test_0009	+0.500	0.286	0.786	14
test_0071	+0.429	0.357	0.786	14
test_0048	+0.417	0.583	1.000	12
test_0079	+0.409	0.364	0.773	22
test_0059	+0.357	0.643	1.000	14
test_0073	+0.350	0.600	0.950	20
test_0012	+0.312	0.438	0.750	16
test_0076	+0.300	0.700	1.000	10
test_0034	+0.300	0.400	0.700	20
Most Hurt				
test_0011	-0.250	0.875	0.625	8
test_0029	-0.200	0.850	0.650	20
test_0020	-0.200	0.600	0.400	10
test_0005	-0.200	0.700	0.500	20
test_0068	-0.167	0.667	0.500	6
test_0022	-0.167	0.722	0.556	18
test_0070	-0.154	0.615	0.462	26
test_0038	-0.147	0.588	0.441	34
test_0014	-0.115	0.692	0.577	26
test_0067	-0.107	0.714	0.607	28

Table 16: **Per-document accuracy change (Δ) for AAEC Task 2.**