Joint Modeling of Entities and Discourse Relations for Coherence Assessment

Wei Liu and Michael Strube
Heidelberg Institute for Theoretical Studies gGmbH
{wei.liu, michael.strube}@h-its.org

Abstract

In linguistics, coherence can be achieved by different means, such as by maintaining reference to the same set of entities across sentences and by establishing discourse relations between them. However, most existing work on coherence modeling focuses exclusively on either entity features or discourse relation features, with little attention given to combining the two. In this study, we explore two methods for jointly modeling entities and discourse relations for coherence assessment. Experiments on three benchmark datasets show that integrating both types of features significantly enhances the performance of coherence models, highlighting the benefits of modeling both simultaneously for coherence evaluation.

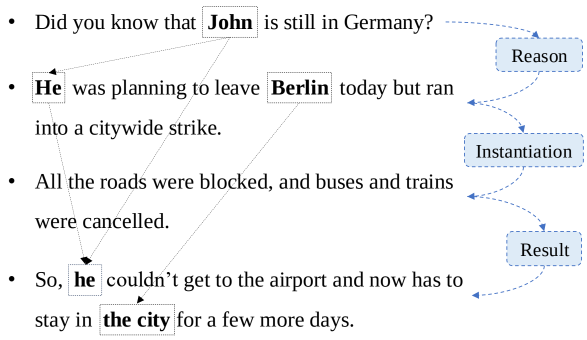

Figure 1: An example of a coherent text, whose coherence should be explained using both entities and discourse relations. We bold the interlinked entities in the text and show the discourse relations between sentences.

1 Introduction

Coherence is a property of well-written texts that makes them easier to read and understand than a sequence of randomly strung sentences (Lapata and Barzilay, 2005). Its modeling benefits many downstream NLP tasks, such as machine translation (Sia and Duh, 2023), topic modeling (Li et al., 2023), text generation (Guan et al., 2023), and dialog generation (Mendonca et al., 2024).

In linguistics, text coherence can be achieved in several ways, with two of the most widely studied being entity-based and discourse relation-based coherence (Reinhart, 1980; Jurafsky and Martin, 2025). Entity-based coherence focuses on how entities are introduced and maintained throughout a text (Prince, 1981; Grosz et al., 1995). In contrast, discourse relation-based coherence considers the logical or rhetorical relationships between sentences (Kehler et al., 2008; Rohde et al., 2018). These perspectives have inspired distinct modeling approaches: entity-based methods (Barzilay and Lapata, 2008; Guinaudeau and Strube, 2013; Tien Nguyen and Joty, 2017; Jeon and Strube, 2022) typically model local coherence by tracking entity transitions, while discourse-based methods (Lin et al., 2011; Feng et al., 2014; Wang et al., 2019; Wu et al., 2023) evaluate coherence based on parsed discourse relations.

While these approaches have proven effective individually, real-world texts often require a more integrated view. In practice, entity and discourse relation cues frequently coexist and interact in complex ways. To illustrate this, we present an example in Figure 1, which contains four sentences and is considered highly coherent. Establishing the coherence using entities is not straightforward in this case, as there are no overlapping entities between the second and third sentences. Instead, we must use a more complex linguistic phenomenon, namely bridging (Clark, 1975; Hou et al., 2018), to link “city” (in “citywide”) and “road”. Meanwhile, the connection between these sentences is more readily explained by a discourse relation (e.g., Instantiation), as the third sentence elaborates on the strike mentioned earlier. However, relying solely on discourse relations also has limitations, as it can compromise the smooth tracking of the protagonist if the referents are unclear. For example, if the final

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21910–21926
November 4-9, 2025

sentence were changed to "So, Maria couldn't get to the airport..." the discourse relation might still hold, but the referent switch (i.e., John → Maria) would disrupt the overall coherence. This underscores the need to jointly consider both entity continuity and discourse structure. Despite their complementary nature, few studies have empirically investigated whether combining these two perspectives leads to more effective coherence assessment.

To address this gap, we propose two approaches for jointly modeling entities and discourse relations in coherence evaluation. The first approach identifies the entities in a document and the discourse relations between sentences, then organizes them, along with the sentences, in a flat structure. We introduce a fusion Transformer that jointly models these elements to assess coherence. The second approach avoids dedicated fusion modules by incorporating entity and discourse relation information directly into prompts, allowing large language models (LLMs) to leverage them during inference.

We evaluate[1] our methods on three benchmarks: two for assessing discourse coherence and one for automatic essay scoring. Our models significantly outperform strong baselines, demonstrating the benefits of joint modeling. Further analysis reveals that integrating both entities and discourse relations enables better learning of coherence patterns, which help to mitigate the effects of imbalanced data distributions in datasets and improve models' generalization across domains.

2 Related Work

Our work is related to existing approaches that enhance coherence modeling using entities, discourse relations, or Transformer-based models.

Entity-based. The most well-known entity-based model is the Entity Grid, proposed by Barzilay and Lapata (2008), which constructs a two-dimensional matrix to capture the transitions of entities between adjacent sentences. This model has been improved by various subsequent efforts, such as incorporating semantically related entities (Filippova and Strube, 2007) and integrating entity-specific features (Elsner and Charniak, 2011). Another prominent entity-centered approach is the Entity Graph, proposed by Guinaudeau and Strube (2013), which measures textual coherence by evaluating the extent to which sentences are connected to each other via shared discourse entities. Building on similar ideas, Mesgar and Strube (2015, 2016) model coherence using the local connectivity structure of sentences. With the rise of deep learning, neural networks have also been applied to capture entity-based coherence patterns. For example, Tien Nguyen and Joty (2017) and Joty et al. (2018) extend the entity grid using convolutional neural networks. Jeon and Strube (2020) introduce a structure-aware model to approximate Centering Theory, which is further refined by Jeon and Strube (2022) through the use of more linguistically grounded units, such as noun phrases and proper names.

Discourse Relation-based. Compared to entity-based models, fewer studies have employed discourse relations for coherence assessment, largely due to the limited performance of early discourse parsers. One of the earliest works in this area is by Lin et al. (2011), who use discourse relations as features for evaluating coherence. Specifically, they adopt an approach similar to the entity grid, constructing a two-dimensional matrix where rows represent sentences and columns represent entities, and each cell (s_i, e_j) contains the set of discourse roles of the entity e_j that appears in the sentence s_i. Feng et al. (2014) extend this approach by replacing shallow discourse relations with deeper ones derived from an RST (Mann and Thompson, 1988) parser. However, Mesgar and Strube (2015) criticize these methods as conceptually flawed, arguing that treating discourse relations as features of entities contradicts their linguistic function, which is to link sentences or elementary discourse units (EDUs). More recently, Wu et al. (2023) propose a multi-task framework that jointly identifies discourse relations between sentences and evaluates the overall coherence of a text.

Unlike these two lines of work focusing solely on entities or discourse relations, we aim to combine both for more effective coherence modeling.

Transformer-based. Our work is also related to recent studies that use Transformer models for coherence assessment. Abhishek et al. (2021) demonstrate that RoBERTa significantly outperforms earlier embedding-based models, with performance further improving under a multi-task training setup incorporating NLI tasks. Laban et al. (2021) use Transformer models to tackle the shuffle test task, achieving near-perfect accuracy (97.88%). To probe the capabilities of language models in coherence prediction, Beyer et al. (2021) design targeted test suites addressing diverse aspects of discourse and dialogue coherence. Building on these

[1] https://github.com/liuwei1206/EntyRelCoh

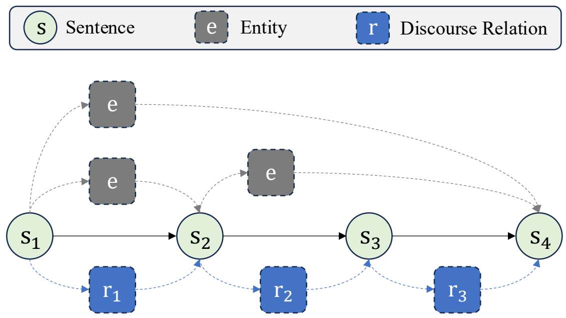

Figure 2: Sentences (in Figure 1) linked by entities and discourse relations.

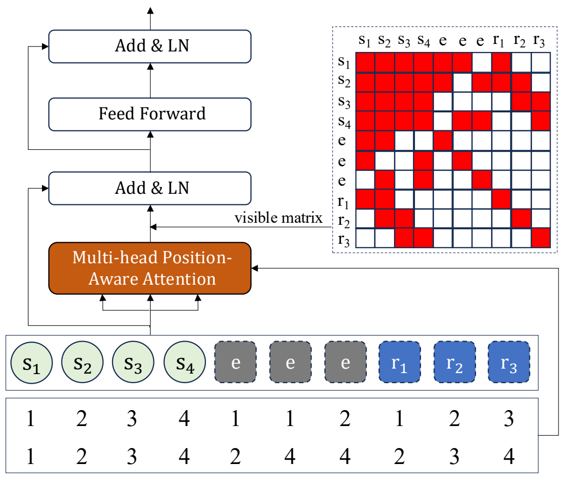

Figure 3: The sentences, entities, and discourse relations in Figure 2 are organized into a flat structure, in which each element is assigned a two-dimensional position, indicating its start and end position in the original sentence sequence. This flat input is then processed by a fusion Transformer.

directions, Zhao et al. (2023) propose DiscoScore, a BERT-based metric inspired by Centering Theory, which models coherence from multiple discourse perspectives and shows a high correlation with human judgments across coherence and factual consistency. More recently, large language models have also been applied to coherence evaluation. Naismith et al. (2023) show that GPT-4 can produce coherence ratings comparable to those of human annotators, accompanied by well-reasoned explanations. Similarly, Mansour et al. (2024) assess ChatGPT and LLaMA on essay scoring tasks, finding that, with appropriate prompting, both models achieve strong performance even in one-shot settings.

3 Method

In this section, we introduce how to identify entities and discourse relations in a document, followed by two methods that use the identified entities and discourse relations to evaluate coherence.

Given a document, we use Stanza (Qi et al., 2020) to identify all nouns and co-references, and to segment the text into sentences. We focus on nouns rather than entities because previous studies have shown that using nouns leads to better performance in coherence modeling (Elsner and Charniak, 2011; Tien Nguyen and Joty, 2017). For discourse relations, we follow prior work(Lin et al., 2011) that adopts the Penn Discourse Treebank (PDTB) framework (Prasad et al., 2008). Specifically, we use the discourse parser `discopy`, developed by Knaebel (2021), to extract relations between adjacent sentences, with a few modifications. First, we use PDTB 3.0 (Webber et al., 2019) instead of PDTB 2.0 (Prasad et al., 2008), as the former includes more relation types and is an improved version of the latter. Second, for implicit discourse relation classification, we use the model proposed by Liu and Strube (2023), which achieves state-of-the-art performance. We provide more details about the parser in Appendix A.

After identifying nouns, coreference relations, and discourse relations, we link two sentences if: (1) they share the same nouns or there is a coreference link between mentions in the sentences, or (2) they are connected by a discourse relation. In the first case, we add an edge labeled "entity" between the sentences, while in the second case, we add an edge labeled with the specific type of discourse relation. Figure 2 shows how the sentences in Figure 1 are linked through the identified entities and discourse relations, forming a graph structure.

However, since the Transformer is designed for sequence modeling (Vaswani et al., 2017), it doesn't naturally handle graph-structured input. One possible solution is to use Graph Neural Networks (GNNs), but standard GNNs are permutation-invariant and cannot capture order information (Wu et al., 2021), which is crucial for coherence modeling (Lapata, 2003). Below we introduce two approaches to address these issues.

3.1 Method I: Fusion

In this approach, we introduce a flat structure to organize sentences, entities, and discourse relations, and design a fusion transformer to jointly model these elements. Figure 3 shows an overview.

In the flat structure, sentences, entities, and discourse relations are concatenated into a sequence.

Each element in this sequence is assigned a two-dimensional position (see the bottom part in Figure 3), indicating its **start** and **end** positions within the original sentence sequence. Take s_1 and r_1 for an example, their positions are $(1, 1)$ and $(1, 2)$, respectively, which means that s_1 is the first sentence in the text and r_1 links the first and second sentences. This flat structure preserves sentence order as well as the connections among sentences, entities, and discourse relations. Its sequential format also makes it well-suited for Transformer models.

To handle this flat structure, we propose a fusion Transformer that enhances the vanilla Transformer with a novel position-aware attention mechanism and a visible matrix. Specifically, we first use a text encoder, such as RoBERTa or LLama, to obtain the representations of sentences, entities, and discourse relations. Then, we input all the elements along with their two-dimensional positions into the position-aware attention. The position-aware attention between the i-th and the j-th elements in the sequence is defined as:

$$\mathbf{A}_{ij} = \mathbf{q}_i\mathbf{k}_j^T + \mathbf{q}_i\mathbf{r}_{i-j}^T + \mathbf{u}\mathbf{k}_j^T + \mathbf{v}\mathbf{r}_{i-j}^T \tag{1}$$

where $\mathbf{q}_i, \mathbf{k}_j, \mathbf{r}_{i-j} = \mathbf{e}_i\mathbf{W}_q, \mathbf{e}_j\mathbf{W}_k, \mathbf{pe}_{i-j}\mathbf{W}_r$, $\mathbf{e}_i$ means the representation of the i-th element, $\mathbf{pe}_{i-j}$ denotes the relative position embedding between the i-th and the j-th elements, and $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_r$, $\mathbf{u}$, $\mathbf{v}$ are trainable parameters. The first and third terms in Eq. 1 are content-based addressing, where the former calculates weight between query and key, and the latter governs a global content bias (Dai et al., 2019). The second and last terms compute weight with relative positional information, which can be used to guide the attention between relevant elements. Since each element in the flat structure has a 2D position, we can calculate four types of relative distances between the i-th and j-th elements: (i) $\mathrm{start_i} - \mathrm{start_j}$; (ii) $\mathrm{start_i} - \mathrm{end_j}$; (iii) $\mathrm{end_i} - \mathrm{start_j}$; (iv) $\mathrm{end_i} - \mathrm{end_j}$. The final relative position embedding between the i-th and j-th elements, i.e., $\mathbf{pe}_{i-j}$, is defined as a non-linear transformation over the four relative distances:

$$\mathbf{pe}_{i-j} = (\mathbf{p}_{s_i - s_j} \otimes \mathbf{p}_{s_i - e_j} \otimes \mathbf{p}_{e_i - e_j} \otimes \mathbf{p}_{e_i - e_j})\mathbf{W}_p \tag{2}$$

The position embedding $\mathbf{p}$ is initialized as in Transformer, where $\mathbf{p}_{pos}^{2k} = \sin\left(pos/10000^{2k/d_{model}}\right)$ and $\mathbf{p}_{pos}^{2k+1} = \cos\left(pos/10000^{2k/d_{model}}\right)$.

To prevent sentences from attending to irrelevant entities and discourse relations, we further introduce a visible matrix $\mathbf{M}$ to guide the attention:

$$\mathbf{M}_{ij} = \begin{cases} 0, & \text{if } \mathrm{C}_1 \mid \mathrm{C}_2 \mid \mathrm{C}_3 \mid \mathrm{C}_4 \\ -\infty, & \text{otherwise} \end{cases} \tag{3}$$

where C_1 is $i = j$ (i.e., self-connection), C_2 is that both i-th and j-th elements are sentences (text content), C_3 is that one element is a sentence and the other is an entity, and the sentence links to the entity (entity patterns), and C_4 is defined as nodes i and j is one sentence and one relation, and the relation works on the sentence (discourse relation patterns). We apply the visible matrix to the attention calculation:

$$\mathbf{A}^* = \mathrm{Softmax}(\mathbf{A} + \mathbf{M}) \tag{4}$$

Then layer normalizations and a feed-forward network (as shown in Figure 3) are applied to produce the text representation. Finally, we input the representation into a softmax classifier, and use the cross-entropy loss for training.

3.2 Method II: Prompt

While the first approach can model coherence using entity and discourse relation information, it relies on an additional fusion module and cannot fully leverage the generative capabilities of Large Language Models (i.e., it merely treats LLMs as a feature extractor). Inspired by Ye et al. (2024), we explore a second approach that uses natural language to describe the connections among sentences, entities, and discourse relations, and then prompts LLMs to take these information into account for coherence assessment. Figure 4 illustrates this approach using the example from Figure 1 and its corresponding connection graph from Figure 2.

Given a graph composed of sentences, entities, discourse relations, and their connections, we traverse all sentence nodes in the order they appear in the text, from left to right. Sentences are added to the prompt and labeled with their position (e.g., s_1, s_2, etc., see Figure 4). For each sentence node, we perform a depth-first search to find all two-hop neighboring nodes that are bridged by an entity or a discourse relation. This allows us to break down the graph into a list of triples, where each triple ($\mathrm{s_i}$, $\mathrm{r_{ij}}$, $\mathrm{s_j}$) includes two sentences, $\mathrm{s_i}$ and $\mathrm{s_j}$, along with the relation $\mathrm{r_{ij}}$ between them. We only retain triples where $\mathrm{i} < \mathrm{j}$, following the natural left-to-right reading order of humans, as suggested by Liu et al. (2023b). For example, the graph in Figure 2 is broken down into the following triples: (s_1, entity, s_2),

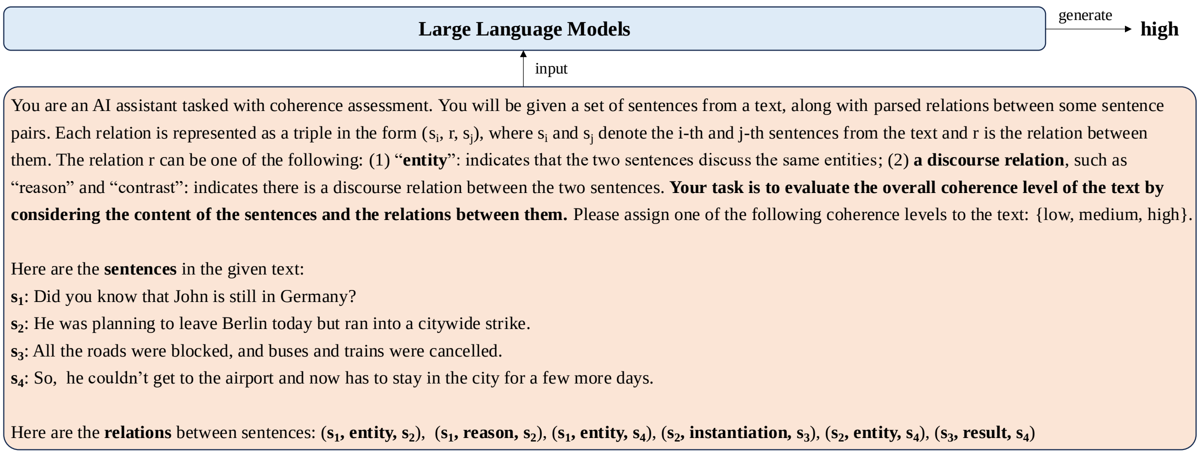

Figure 4: Illustration of our second approach. We use natural language to describe the relationships between sentences, entities, and discourse relations in Figure 2, presenting the graph structure in a concise and intuitive way. We then instruct LLMs to consider these elements for coherence assessment.

(s_1, reason, s_2), (s_1, entity, s_4), (s_2, instantiation, s_3), (s_2, entity, s_4), (s_3, result, s_4). These triples are expressed in natural language format, making them easy for LLMs to process. More importantly, they retain all the connection information between sentences, entities, and discourse relations. Finally, we include the list of triples in the prompt and instruct the LLMs to assess coherence by considering both the content of the sentences and the patterns of entities and discourse relations between them (see Figure 4).

4 Experiments

Datasets. We conduct experiments on three widely used corpora in coherence modeling: GCDC (Lai and Tetreault, 2018), CoheSentia (Maimon and Tsarfaty, 2023), and TOEFL (Blanchard et al., 2013). GCDC is a corpus designed for evaluating discourse coherence, containing texts from four distinct domains: **Yahoo** online forum posts, **Enron** emails, emails from Hillary **Clinton**'s office, and **Yelp** business reviews. Each text in the dataset is rated by experts on a scale of 1 to 3, indicating low, medium, and high levels of coherence. CoheSentia is another dataset used to assess discourse coherence. Unlike GCDC, which consists of real-world texts, CoheSentia contains stories generated by GPT-3 and is annotated by humans with coherence scores ranging from 1 to 5. However, the score distribution is highly imbalanced,[2] which makes it difficult for models to converge during training (Maimon and Tsarfaty, 2023). To address this, we group scores 1 and 2 as low coherence, scores 3 and 4 as medium coherence, and score 5 as high coherence. The TOEFL dataset was originally created for automated essay scoring but has since been widely used to evaluate coherence models (Burstein et al., 2010; Jeon and Strube, 2020). It includes essays written in response to eight prompts (P1 to P8) along with score levels (low/medium/high) for each essay.

Implementation Details. We implement our models using the PyTorch library. For Method I, we experiment with two widely used text encoders (Abhishek et al., 2021; Parmar et al., 2024): the pre-trained language model $\mathrm{RoBERTa_{base}}$ (Liu et al., 2019b) and the large language model Llama-3.1-8B-Instruction (Grattafiori et al., 2024).[3] Training is performed using the AdamW optimizer with an initial learning rate of 1e-3, a batch size of 32, and a maximum of 20 epochs.

For Method II, which is specifically designed for large language models (LLMs), we evaluate it using Llama-3.1-8B-Instruction.[3] The evaluation is conducted under two settings: **zero-shot** and **fine-tuned**. In the **zero-shot** setting, the model is not trained beforehand; instead, it is directly prompted to generate labels. This setup tests whether incorporating entity and discourse relation features can help with coherence evaluation in cold-start scenarios. In the **fine-tuned** setting,

[2]Over 50% of the data is labeled with a score of 5.

[3]We use the 8B LLaMA model instead of the 70B due to memory limitations that prevent fine-tuning larger models. However, our resources do support zero-shot experiments with the 70B model. To maintain consistency across settings, we use the 8B model throughout the main text, but include zero-shot results for the 70B model in the Appendix E.

Model			GCDC					CoheSentia
			Clinton	Enron	Yahoo	Yelp	Avg	
Jeon and Strube (2022)			$64.20_{0.4}$	$55.30_{0.3}$	$58.40_{0.2}$	$57.30_{0.2}$	58.90	-
Liu et al. (2023b)			$66.20_{0.8}$	$57.00_{0.8}$	$63.65_{0.7}$	$58.05_{1.2}$	61.23	-
Fusion	RoBERTa	TextOnly	$64.55_{0.7}$	$57.50_{0.9}$	$60.05_{0.4}$	$58.20_{0.8}$	60.10	$60.64_{1.5}$
		TextEnty	$66.20_{0.8}$	$58.80_{1.1}$	$63.15_{0.9}$	$59.20_{1.1}$	61.83	$63.13_{2.0}$
		TextRel	$66.45_{0.9}$	$59.70_{1.0}$	$63.35_{1.1}$	$60.40_{1.3}$	62.48	$63.74_{1.8}$
		Our Method I	$\mathbf{67.60}_{0.5}$	$60.50_{0.3}$	$63.75_{0.5}$	$61.10_{0.4}$	**63.24**	$66.24_{1.6}$
	Llama	TextOnly	$63.55_{0.5}$	$56.65_{0.8}$	$59.45_{0.8}$	$57.45_{1.0}$	59.27	$63.13_{1.2}$
		TextEnty	$64.80_{0.8}$	$58.10_{0.4}$	$62.10_{0.5}$	$57.90_{0.8}$	60.73	$65.80_{1.5}$
		TextRel	$65.10_{0.7}$	$58.75_{0.4}$	$62.85_{0.3}$	$59.35_{0.5}$	61.51	$66.65_{1.6}$
		Our Method I	$67.25_{0.4}$	$60.10_{0.3}$	$\mathbf{64.10}_{0.5}$	$\mathbf{61.30}_{0.5}$	63.18	$\mathbf{69.12}_{1.5}$
Prompt	Llama zero-shot	TextOnly	54.50	38.00	34.00	40.50	40.88	50.10
		TextEnty	55.00	39.00	41.50	44.50	45.00	51.35
		TextRel	57.50	41.00	42.00	45.50	46.50	52.17
		Our Method II	56.50	41.00	42.00	48.00	46.88	53.83
	Llama fine-tuned	TextOnly	$63.55_{0.8}$	$56.80_{0.9}$	$60.05_{1.0}$	$55.45_{1.2}$	58.96	$64.95_{1.4}$
		TextEnty	$65.00_{1.2}$	$57.60_{0.5}$	$60.45_{1.0}$	$56.30_{0.9}$	59.84	$65.38_{1.5}$
		TextRel	$64.55_{0.7}$	$59.10_{0.5}$	$61.10_{0.7}$	$57.25_{0.5}$	60.50	$66.42_{1.4}$
		Our Method II	$65.15_{0.6}$	$\mathbf{60.55}_{1.2}$	$62.05_{1.2}$	$57.55_{0.5}$	61.33	$67.28_{1.1}$

Table 1: Mean accuracy results (with std) on GCDC and CoheSentia.

we fine-tune the Llama model using LoRA for 3 epochs, with a learning rate of 5e-5 and a batch size of 2. This setup evaluates whether instruction-tuning the LLM to consider entities and discourse relations can enhance its performance.

To account for training variability, we perform 10-fold cross-validation on the GCDC training dataset (Lai and Tetreault, 2018), 5-fold cross-validation on the CoheSentia corpus, and 5-fold cross-validation on the dataset for each prompt in the TOEFL corpus (Taghipour and Ng, 2016). Following prior work, we use standard accuracy (Acc, %) as our primary evaluation metric.[4]

Baselines. To validate the importance of modeling entities and discourse relations simultaneously, we compare it with the following baselines:

- **TextOnly**. This baseline relies solely on textual information for coherence modeling. In Method I, this involves using a text encoder to obtain sentence representations, a sentence-level transformer to capture coherence patterns, and a softmax classifier for prediction. In Method II, it prompts LLMs to evaluate coherence based only on the text.

- **TextEnty**. This is an ablated version of our approach in which the discourse relation elements are removed from the sentence-entity-discourse relation graph.

- **TextRel**. This is another ablated version of our method, where we remove the entity elements from the graph.

Further, we compare our approaches against previous state-of-the-art models on each corpus. For more details on the datasets, implementation, and baselines, please refer to Appendix B.

4.1 Overall Results

GCDC / CoheSentia. Table 1 shows the results on GCDC and CoheSentia datasets, where the "Fusion" block shows the results relying on an extra fusion module to integrate entity and discourse relation features, while the "Prompt" block presents the results using natural languages to incorporate entity and discourse relation patterns into the input prompt of LLMs.

For the Fusion style, we show the results based on RoBERTa and LLama. Regardless of whether RoBERTa or Llama is used as the text encoder, TextEnty and TextRel consistently outperform the TextOnly baseline on GCDC and CoheSentia. This

[4]We also report the results of Macro-F1 in Appendix C.

Model			P1	P2	P3	P4	P5	P6	P7	P8	Avg
Jeon and Strube (2022)			78.38	75.70	76.58	76.56	79.10	76.41	75.03	74.54	76.54
Liu et al. (2023b)			$75.79_{1.1}$	$76.25_{1.1}$	$74.14_{1.2}$	$75.81_{0.7}$	$77.01_{0.9}$	$77.08_{1.1}$	$73.55_{0.8}$	$72.91_{0.7}$	75.34
Fusion	RoBERTa	TextOnly	$76.36_{0.9}$	$75.10_{1.0}$	$75.29_{0.5}$	$75.33_{1.5}$	$75.90_{1.0}$	$75.61_{1.9}$	$73.76_{0.9}$	$73.34_{1.1}$	75.08
		TextEnty	$79.05_{1.4}$	$77.15_{1.2}$	$77.73_{0.8}$	$76.98_{1.3}$	$77.64_{1.6}$	$78.32_{1.5}$	$76.49_{1.3}$	$75.79_{1.0}$	77.39
		TextRel	$78.94_{0.8}$	$77.41_{0.7}$	$77.80_{0.8}$	$77.55_{0.8}$	$78.49_{0.9}$	$78.33_{1.5}$	$77.08_{1.2}$	$76.25_{0.5}$	77.73
		Our Method I	$79.92_{0.8}$	$\mathbf{78.46}_{0.9}$	$\mathbf{78.68}_{0.9}$	$\mathbf{78.25}_{1.2}$	$79.23_{1.1}$	$79.42_{1.27}$	$78.21_{0.9}$	$77.13_{1.1}$	**78.66**
	Llama	TextOnly	$75.17_{0.8}$	$73.88_{1.3}$	$73.63_{1.6}$	$73.67_{1.4}$	$75.89_{1.0}$	$75.10_{0.9}$	$73.67_{1.4}$	$72.87_{1.5}$	74.24
		TextEnty	$77.03_{0.8}$	$75.59_{1.4}$	$75.14_{1.5}$	$75.20_{1.5}$	$77.07_{0.9}$	$77.12_{0.8}$	$75.48_{0.6}$	$74.17_{1.4}$	75.85
		TextRel	$76.35_{0.9}$	$76.40_{0.7}$	$75.98_{0.5}$	$75.40_{1.2}$	$76.64_{1.7}$	$76.65_{1.6}$	$75.18_{1.1}$	$75.16_{1.3}$	75.97
		Our Method I	$78.24_{1.7}$	$78.11_{1.9}$	$77.01_{1.1}$	$76.59_{1.1}$	$79.23_{1.3}$	$\mathbf{79.47}_{1.6}$	$77.32_{1.1}$	$76.50_{1.8}$	77.81
Prompt	Llama zero-shot	TextOnly	51.39	55.19	52.72	50.63	54.37	50.62	46.92	49.44	51.41
		TextEnty	56.85	53.78	54.48	54.00	53.83	57.15	55.89	54.64	55.08
		TextRel	58.51	56.45	54.73	55.59	56.43	57.19	57.41	53.72	56.25
		Our Method II	59.90	57.75	56.73	56.13	57.28	58.02	58.19	55.91	57.49
	Llama fine-tuned	TextOnly	$79.03_{1.1}$	$76.76_{1.4}$	$76.24_{1.5}$	$77.52_{1.4}$	$79.49_{1.4}$	$76.02_{1.4}$	$76.69_{1.1}$	$75.28_{0.9}$	77.13
		TextEnty	$\mathbf{80.13}_{1.2}$	$76.63_{1.2}$	$75.64_{1.3}$	$77.73_{1.0}$	$79.55_{1.5}$	$76.57_{1.6}$	$\mathbf{78.95}_{1.4}$	$76.41_{1.3}$	77.70
		TextRel	$79.35_{1.5}$	$77.15_{1.6}$	$77.16_{1.4}$	$76.61_{1.2}$	$80.15_{1.1}$	$75.41_{1.5}$	$78.29_{1.3}$	$76.89_{1.4}$	77.63
		Our Method II	$80.02_{1.6}$	$77.92_{1.5}$	$77.58_{1.2}$	$78.13_{1.3}$	$\mathbf{81.13}_{1.5}$	$77.29_{1.3}$	$77.88_{1.0}$	$\mathbf{77.18}_{1.5}$	78.39

Table 2: Mean accuracy results (with std) on TOEFL dataset.

suggests that incorporating entity or discourse relation features enhances coherence assessment, which is in line with the findings of previous entity-based (Jeon and Strube, 2022) and discourse relation-based studies (Wu et al., 2023). The improvement of TextRel over TextOnly is greater than that of TextEnty over TextOnly. This is because, in both GCDC and CoheSentia, discourse relations are more commonly used to connect sentences than entity cues. For instance, discourse relations like cause and concession are frequently employed in CoheSentia to make stories more compact and engaging (Chaturvedi et al., 2017). Our Method I significantly outperforms both the TextEnty and TextRel baselines, showing a 1% to 2% improvement on GCDC and approximately a 3% gain on CoheSentia. These results highlight the value of jointly modeling entity and discourse relation features for effective coherence assessment.

For the Prompt style, we present the results of Llama in both zero-shot and fine-tuned settings. In the zero-shot setting, incorporating entity and discourse relation information enhances Llama's performance in coherence assessment. On GCDC, TextEnty and TextRel outperform the TextOnly baseline by more than 4% to 5%. In contrast, the improvement on CoheSentia is more modest, with gains of about 1% to 2%. Combining these features further boosts performance, leading to improvements of over 6 points on GCDC and 3.5% on CoheSentia, compared to the TextOnly baseline. These results suggest that prior knowledge of entity- and discourse relation-based coherence can be effectively leveraged for coherence assessment in cold-start scenarios. When fine-tuning LLaMA with LoRA, the performance improvements of TextEnty, TextRel, and EntyRel over TextOnly still exists, but the gains are smaller compared to the zero-shot setting. We speculate that this is because fine-tuning allows the model to somewhat implicitly capture coherence-relevant signals, such as entity transition and discourse relations (Xiao et al., 2021), so the explicit incorporation of them leads to limited improvement.

TOEFL. Results on TOEFL are shown in Table 2. Similar to the findings on GCDC and CoheSentia, both entity and discourse relation patterns contribute positively to the task in the fusion setting. Specifically, TextEnty and TextRel outperform the TextOnly baseline by 2% to 3% when using RoBERTa or Llama as the text encoder. Combining entity and discourse relation features further enhances performance. Our Method I using RoBERTa as the text encoder achieves an average accuracy of 78.66%, significantly outperforming the previous state-of-the-art model (Jeon and Strube, 2022). We observe similar results in the prompt setting: in the zero-shot scenario, Method II achieves an accuracy of 6.08% higher than the TextOnly baseline, and 1.26% higher in the fine-tuned setting.

5 Analysis

To understand how jointly modeling entities and discourse relations contributes to coherence assessment, we analyze the accuracy of different models

[There's two older men here that are always in a lousy mood.]$_1$ [They can't wait for 62.]$_2$ [Andrea M. is correct about the cart corrals.]$_3$ …[Really saves us money in that dept.]$_7$ [But I can only give 3 stars on account that their pricing for vegetables and other items is a ripoff.]$_8$ [Why is there only one dollar sign showing above?]$_9$ [People are SOOOO brainwashed into thinking that Fry's and Walmart save you money.]$_{10}$ …

(s_1, cause, s_2), (s_2, norel, s_3), (s_3, level-of-detail, s_4), (s_4, norel, s_5), (s_5, norel, s_6), (s_6, cause, s_7), (s_7, norel, s_8), (s_7, entity, s_{10}), …

Model	**Prediction**	**Explanation**
TextOnly	High	…… The reviewer's tone is conversational and informal, but their points are clearly stated and easy to follow.
TextEnty	Low	The relation between s_7 and s_{10} is "entity", indicating that they discuss the same entities. However, s_7 is discussing the cost-effectiveness of prescriptions, while s_{10} is discussing the cost-effectiveness of Fry's and Walmart. This lack of coherence between the entities being discussed suggests a low coherence level.
TextRel	Medium	The relations between sentences provide a clear narrative flow, but the contrast between the customer's satisfaction with the store's prices and their dissatisfaction with the pricing of vegetables creates a sense of inconsistency.
Ours	Low	The relations between sentences provide a clear narrative flow, but the content of the sentences are often abrupt and lack cohesion, indicating a low level of coherence.

[Look at you...such a hoax spreader...ha]$_1$ [Anyway, i was so busy yesterday i didn't have a chance to do anything about it anyway.]$_2$ [so no harm, no foul.]$_3$ [how are you doing?]$_4$ [as for Enron, as you well know our stock's not doing so great these days, but i'm not overly stressed.]$_5$ … [Can't wait to hear how your preparations are coming along...for the wedding and the move.]$_7$ [btw, have you set a firm date yet?]$_8$

(s_1, norel, s_2), (s_2, cause, s_3), (s_3, norel, s_4), (s_4, conjunction, s_5), (s_5, conjunction, s_6), (s_5, entity, s_6), (s_6, conjunction, s_7), (s_7, conjunction, s_8)

Model	**Prediction**	**Explanation**
TextOnly	Medium	… seemingly unrelated statements, …. the presence of a few cohesive elements, … indicates some level of coherence.
TextEnty	Medium	… various unrelated topics… s5 and s6 discuss the same entities, which is a positive sign of coherence.
TextRel	Low	The text discusses various unrelated topics … The relations between sentences are mostly conjunctions and sometimes norel, which indicate a simple sequence of ideas rather than a deeper connection between them.
Ours	Low	The text discusses various unrelated topics… The relations between sentences are mostly conjunctions, indicating a lack of strong logical connections between the ideas.

Figure 5: Two examples (truncated) showing how entities and discourse relations aid coherence assessment. Both texts are labeled as low coherence. We use a zero-shot prompt setting, and the "explanation" refers to Llama's brief justification for its prediction.

across each coherence label. Table 3 presents the models' performance on the TOEFL P5 dataset in both the Fusion setting (with Llama as the text encoder) and the fine-tuned Prompt setting. TextOnly exhibits a strong bias, achieving high accuracy on "medium" and "high" coherence labels but significantly lower accuracy on the "low" label. We attribute this to the imbalanced label distribution in the TOEFL P5 dataset, where over 90% of samples are annotated as "medium" or "high" coherence. TextEnty and TextRel help mitigate this bias by incorporating entity and discourse relation information, respectively. For example, in the Fusion setting, they improve accuracy on low-coherence data by 6.57% and 7.69%. Our Methods I and II go further by jointly modeling entities and discourse relations, resulting in the smallest performance gap across all three coherence levels. These results suggest that incorporating entities and discourse relations helps the model learn more effective coherence patterns and improves its robustness to imbalanced data distributions.

To better understand how entities and discourse relations influence model behavior, we present two case studies in Figure 5. The two examples are

		Low	Medium	High	Range
Fusion (Llama)	TextOnly	66.67	78.99	77.88	12.32
	TextEnty	73.24	80.44	76.79	7.20
	TextRel	74.36	80.45	78.41	6.09
	Our Method I	81.16	81.99	77.19	4.80
Prompt (fine-tuned)	TextOnly	68.22	83.29	82.93	15.07
	TextEnty	71.70	85.23	85.49	13.79
	TextRel	70.59	84.09	84.05	13.50
	Our Method II	73.47	85.39	84.71	11.92

Table 3: Accuracy results for each coherence label on TOEFL P5. Range indicates the difference between the highest and lowest values.

from GCDC corpus and annotated as low coherence. In both cases, we use a zero-shot prompt setting, asking Llama to evaluate the coherence level of a given text and provide a brief explanation for its assessment (see Appendix D for details). As shown in the first example, without entity and discourse relation information (i.e., TextOnly), Llama evaluates the text as having high coherence. TextRel identifies some inconsistencies but still fails to classify it as medium coherence. In contrast, TextEnty and Our Method II correctly assess the text as having low coherence, due to the lack of cohesion, specifically, missing entity-based signals. In the second example, all models recognize

		Enron → Others	TOEFL P1 → Others
Fusion (Llama)	TextOnly	47.48	68.79
	TextEnty	50.62 (+3.14)	72.02 (+3.23)
	TextRel	50.98 (+3.55)	72.87 (+4.08)
	Our Method I	53.82 (+6.34)	74.40 (+5.61)
Prompt (fine-tuned)	TextOnly	52.50	76.72
	TextEnty	53.67 (+1.17)	78.42 (+1.70)
	TextRel	54.75 (+2.25)	78.15 (+1.43)
	Our Method II	56.00 (+3.50)	78.60 (+1.88)

Table 4: Accuracy of models in a cross-domain setting.

that the sentences in the text cover various unrelated topics. However, TextOnly and TextEnty are slightly influenced by the presence of cohesive elements, leading them to predict the text as medium coherence. In contrast, TextRel and Our Method II correctly and confidently classify it as low coherence, due to the lack of logical connections between the sentences. These two cases effectively illustrate the importance of modeling both entity and discourse relation patterns for accurate coherence assessment.

To assess whether our models have truly learned more robust coherence patterns, we further evaluate their transferability in cross-domain settings. Specifically, we train TextOnly, TextEnty, TextRel, and Our Method in both Fusion and Prompt settings on the Enron subset of GCDC (or Prompt 5 of TOEFL) and test their performance on other subsets of GCDC (or other TOEFL prompts). Table 4 presents the results. Both TextEnty and TextRel consistently outperform the TextOnly baseline in cross-domain settings, indicating that entity and discourse relation patterns are effective domain-agnostic features for coherence assessment. Moreover, our methods achieve the best performance across all cross-domain experiments, demonstrating the effectiveness of jointly modeling entities and discourse relations.

6 Conclusions

This paper explores whether combining entity and discourse relation information improves coherence modeling. We propose two novel methods that jointly model entities and discourse relations for coherence assessment. Experiments on three benchmark datasets show that our approaches consistently outperform strong baselines, emphasizing the value of integrating both features. Additionally, we demonstrate that these features enhance model robustness in scenarios with imbalanced labels and across different domains.

Limitations

Our work has several limitations. First, the PDTB parser used in this study is far from perfect. Future research should focus on developing more powerful parsers to support discourse relation analysis for coherence modeling. For instance, it would be worthwhile to explore whether LLM-based approaches can produce better PDTB parsing results. Second, our experiments are limited to PDTB-style discourse relations. Extending the analysis to other frameworks, such as RST (Mann and Thompson, 1988), could offer valuable insights. Finally, due to budget and computational constraints, we only experimented with Llama-8B (and only used Llama-70B in zero-shot setting). It would be interesting to evaluate our approach using other or larger language models, such as GPT-4.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their comments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany.

References

Tushar Abhishek, Daksh Rawat, Manish Gupta, and Vasudeva Varma. 2021. Transformer models for text coherence assessment. *arXiv preprint arXiv:2109.02176*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? targeted evaluation of coherence prediction from language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 681–684, Los Angeles, California. Association for Computational Linguistics.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1603–1614, Copenhagen, Denmark. Association for Computational Linguistics.

Herbert H. Clark. 1975. Bridging. In *Theoretical Issues in Natural Language Processing*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129, Portland, Oregon, USA. Association for Computational Linguistics.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Katja Filippova and Michael Strube. 2007. Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 139–142, Saarbrücken, Germany. DFKI GmbH.

Xiyan Fu and Anette Frank. 2023. SETI: Systematicity evaluation of textual inference. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4101–4114, Toronto, Canada. Association for Computational Linguistics.

Xiyan Fu and Anette Frank. 2024a. Compositional structured explanation generation with dynamic modularized reasoning. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 385–401, Mexico City, Mexico. Association for Computational Linguistics.

Xiyan Fu and Anette Frank. 2024b. Exploring continual learning of compositional generalization in NLI. *Transactions of the Association for Computational Linguistics*, 12:912–932.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Jian Guan, Zhenyu Yang, Rongsheng Zhang, Zhipeng Hu, and Minlie Huang. 2023. Generating coherent narratives by learning dynamic and discrete entity states with a contrastive framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12836–12844.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria. Association for Computational Linguistics.

Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

Sungho Jeon and Michael Strube. 2020. Centering-based neural coherence modeling with hierarchical discourse segments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7458–7472, Online. Association for Computational Linguistics.

Sungho Jeon and Michael Strube. 2022. Entity-based neural local coherence modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7787–7805, Dublin, Ireland. Association for Computational Linguistics.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Shafiq Joty, Muhammad Tasnim Mohiuddin, and Dat Tien Nguyen. 2018. Coherence modeling of asynchronous conversations: A neural entity grid approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 558–568, Melbourne, Australia. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L. Elman. 2008. Coherence and coreference revisited. *J. Semant.*, 25(1):1–44.

René Knaebel. 2021. discopy: A neural system for shallow discourse parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A. Hearst. 2021. Can transformer models measure coherence in text: Re-thinking the shuffle test. 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)*, pages 1058–1064, Online. Association for Computational Linguistics.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. Association for Computational Linguistics.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 545–552, Sapporo, Japan. Association for Computational Linguistics.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, page 1085–1090, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Raymond Li, Felipe Gonzalez-Pizarro, Linzi Xing, Gabriel Murray, and Giuseppe Carenini. 2023. Diversity-aware coherence loss for improving neural topic models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1710–1722, Toronto, Canada. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics.

Wei Liu, Yi Fan, and Michael Strube. 2023a. HITS at DISRPT 2023: Discourse segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49, Toronto, Canada. The Association for Computational Linguistics.

Wei Liu, Xiyan Fu, and Michael Strube. 2023b. Modeling structural similarities between documents for coherence assessment with graph convolutional networks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7792–7808, Toronto, Canada. Association for Computational Linguistics.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021. Lexicon enhanced Chinese sequence labeling using BERT adapter. 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)*, pages 5847–5858, Online. Association for Computational Linguistics.

Wei Liu and Michael Strube. 2023. Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712, Toronto, Canada. Association for Computational Linguistics.

Wei Liu and Michael Strube. 2025. Discourse relation-enhanced neural coherence modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4748–4762, Vienna, Austria. Association for Computational Linguistics.

Wei Liu, Stephen Wan, and Michael Strube. 2024. What causes the failure of explicit to implicit discourse relation recognition? 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 2738–2753, Mexico City, Mexico. Association for Computational Linguistics.

Wei Liu, Tongge Xu, Qinghua Xu, Jiayu Song, and Yueran Zu. 2019a. An encoding strategy based word-character LSTM for Chinese NER. 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 2379–2389, Minneapolis, Minnesota. 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. 2019b. RoBERTa: A robustly optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Aviya Maimon and Reut Tsarfaty. 2023. COHESENTIA: A novel benchmark of incremental versus holistic assessment of coherence in generated texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5328–5343, Singapore. Association for Computational Linguistics.

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

Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786, Torino, Italia. ELRA and ICCL.

John Mendonca, Isabel Trancoso, and Alon Lavie. 2024. ECoh: Turn-level coherence evaluation for multilingual dialogues. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 516–532, Kyoto, Japan. Association for Computational Linguistics.

Mohsen Mesgar and Michael Strube. 2015. Graph-based coherence modeling for assessing readability. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 309–318, Denver, Colorado. Association for Computational Linguistics.

Mohsen Mesgar and Michael Strube. 2016. Lexical coherence graph modeling using word embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1414–1423, San Diego, California. Association for Computational Linguistics.

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 68–82, Online. Association for Computational Linguistics.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

Juri Opitz and Sebastian Burst. 2019. Macro F1 and macro F1. *CoRR*, abs/1911.03347.

Mihir Parmar, Hanieh Deilamsalehy, Franck Dernoncourt, Seunghyun Yoon, Ryan A. Rossi, and Trung Bui. 2024. Towards enhancing coherence in extractive summarization: Dataset and experiments with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19810–19820, Miami, Florida, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Tanya Reinhart. 1980. Conditions for text coherence. *Poetics Today*, 1(4):161–180.

Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267, Melbourne, Australia. Association for Computational Linguistics.

Suzanna Sia and Kevin Duh. 2023. In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 173–185, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Dat Tien Nguyen and Shafiq Joty. 2017. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1320–1330, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xinhao Wang, Binod Gyawali, James V. Bruno, Hillary R. Molloy, Keelan Evanini, and Klaus Zechner. 2019. Using Rhetorical Structure Theory to assess discourse coherence for non-native spontaneous speech. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 153–162, Minneapolis, MN. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse TreeBank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Hongyi Wu, Xinshu Shen, Man Lan, Shaoguang Mao, Xiaopeng Bai, and Yuanbin Wu. 2023. A multi-task dataset for assessing discourse coherence in Chinese essays: Structure, theme, and logic analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6673–6688, Singapore. Association for Computational Linguistics.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.

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*, pages 4139–4152, Online. Association for Computational Linguistics.

Ruosong Ye, Caiqi Zhang, Runhui Wang, Shuyuan Xu, and Yongfeng Zhang. 2024. Language is all a graph needs. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1955–1973, St. Julian's, Malta. Association for Computational Linguistics.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Explicit	Distribution	Implicit	Distribution
Asynchronous	8.69%	Asynchronous	4.64%
Cause	7.87%	Cause	24.23%
Concession	19.94%	Cause+Belief	0.82%
Condition	5.99%	Concession	6.72%
Conjunction	36.55%	Condition	0.85%
Contrast	4.58%	Conjunction	20.84%
Disjunction	1.23%	Contrast	3.86%
Instantiation	1.30%	Equivalence	1.21%
Level-of-detail	1.01%	Instantiation	6.84%
Manner	1.23%	Level-of-detail	14.60%
Negative-condition	0.54%	Manner	0.74%
Purpose	1.63%	Purpose	3.31%
Similarity	0.42%	Substitution	1.34%
Substitution	0.96%	Synchronous	2.35%
Synchronous	8.07%	NoRel	8.18%

Table 5: Explicit and Implicit relations used in this study and their distribution in the training corpus.

Dataset		Split	#Doc	Avg #Sent	Avg #Word
GCDC	Clinton	Train	1000	8.9	182.9
		Test	200	8.8	186.0
	Enron	Train	1000	9.2	185.1
		Test	200	9.3	191.1
	Yahoo	Train	1000	7.8	157.2
		Test	200	7.8	162.7
	Yelp	Train	1000	10.4	178.2
		Test	200	10.1	179.1
CoheSentia	-	Total	483	7.0	122.2
TOEFL	Prompt 1	Total	1656	13.7	339.1
	Prompt 2	Total	1562	15.7	357.8
	Prompt 3	Total	1396	14.7	343.5
	Prompt 4	Total	1509	15.1	338.0
	Prompt 5	Total	1648	15.2	358.4
	Prompt 6	Total	960	15.3	358.3
	Prompt 7	Total	1686	14.0	336.6
	Prompt 8	Total	1683	14.7	340.9

Table 6: Statistics of datasets, where #Doc, #Sent, and #Word mean the number of documents, sentences, and words, respectively.

A PDTB Parser

We use an updated version of discopy (Knaebel, 2021) to parse discourse relations in documents. The first update involves replacing the PDTB 2.0 (Prasad et al., 2008) relation set with PDTB 3.0 (Webber et al., 2019). Specifically, we focus on identifying both explicit and implicit discourse relations between adjacent sentences. For explicit relations, we select 15 types that have sufficient training data (Liu et al., 2023a, 2024). For implicit relations, we include the 14 most frequent types, along with a "NoRel" label to account for cases where no relation is present—common in low-coherence texts. Table 5 lists all the relations used in this study along with their distribution in PDTB 3.0.

The second update incorporates the model proposed by Liu and Strube (2023) for recognizing implicit relations, due to its state-of-the-art performance. We implement the parser using RoBERTa and train it on PDTB 3.0, following the data split introduced by Ji and Eisenstein (2015). The parser achieves 89.61% accuracy on the explicit test set and 67.80% on the implicit test set of PDTB 3.0.

B Experimental Settings

B.1 Dataset

The GCDC dataset includes texts from four domains: online forum posts from Yahoo, emails from the Enron corpus, emails from Hillary Clinton's office, and online business reviews from Yelp. The CoheSentia datasets consists of stories generated by GPT-3. The TOEFL dataset comprises essays written in response to eight different prompts. Table 6 presents statistics for these three corpora.

B.2 Implementation

Fusion. In the Fusion setting, we use a text encoder, such as RoBERTa or LLaMA, to obtain sentence representations. This is done by passing a sentence through the encoder, extracting token-level representations, and then averaging the representations of the tokens within the sentence. We experimented with both average pooling and [CLS] pooling methods. Our results show that average pooling consistently outperforms [CLS] pooling (Liu and Strube, 2025). For instance, on the TOEFL P1 dataset using a RoBERTa encoder, the accuracy of the TextOnly baseline and Our Method I with average pooling is 76.36 and 80.55, respectively, compared to 72.58 and 77.56 with [CLS] pooling. This improvement is likely because average pooling incorporates information from all tokens in the sentence, preserving more linguistic features. In contrast, [CLS] pooling relies solely on the [CLS] token's representation, which can result in the loss of important information. Similar results are observed for average pooling and [CLS] pooling in Mosbach et al. (2020). For entity and discourse relation elements in the flat structure, we convert them as vectors using GloVe embeddings (Pennington et al., 2014). We use two layers of Fusion Transformers to jointly model sentences, entities, and discourse relations. Each layer consists of 8 attention heads and has a hidden size of 256. The model is trained using the AdamW optimizer with an initial learning rate of 1e-3, a batch size of 32, a dropout rate of 0.1, and a maximum of 20 training epochs.

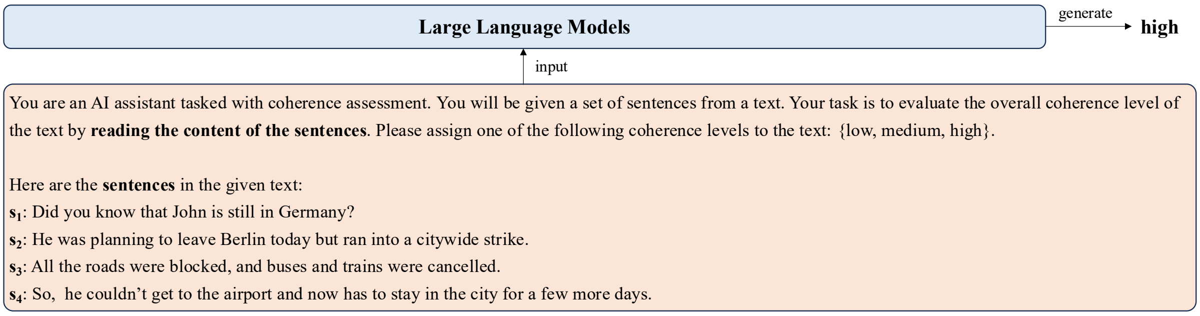

Figure 6: Illustration of TextOnly baseline in the Prompt setting. We instruct LLMs to consider only textual content for coherence assessment.

Model			GCDC					CoheSentia
			Clinton	Enron	Yahoo	Yelp	Avg	
Fusion	RoBERTa	TextOnly	47.58$_{0.9}$	48.74$_{1.0}$	45.71$_{0.9}$	45.63$_{0.8}$	46.92	57.08$_{1.7}$
		TextEnty	52.38$_{1.2}$	48.84$_{1.4}$	48.21$_{1.8}$	47.24$_{1.6}$	49.17	59.94$_{2.1}$
		TextRel	52.42$_{1.3}$	51.04$_{1.5}$	48.56$_{1.7}$	47.35$_{1.8}$	49.84	60.35$_{1.9}$
		Our Method I	**54.49**$_{1.6}$	51.27$_{1.1}$	48.63$_{0.8}$	47.86$_{1.1}$	50.56	62.98$_{1.7}$
	Llama	TextOnly	47.54$_{1.8}$	48.73$_{1.6}$	44.38$_{1.0}$	46.09$_{1.4}$	46.68	59.95$_{1.6}$
		TextEnty	50.82$_{1.0}$	50.98$_{1.1}$	47.74$_{0.8}$	47.29$_{1.4}$	49.20	62.52$_{2.0}$
		TextRel	49.73$_{1.7}$	50.77$_{1.6}$	47.37$_{0.9}$	**48.53**$_{0.6}$	49.10	63.67$_{2.1}$
		Our Method I	53.78$_{1.3}$	**52.37**$_{1.6}$	**50.50**$_{1.3}$	47.59$_{1.3}$	**51.06**	**65.25**$_{1.8}$
Prompt	Llama zero-shot	TextOnly	34.78	32.02	32.39	32.79	33.88	40.06
		TextEnty	40.24	34.71	38.69	36.56	37.55	41.09
		TextRel	41.43	36.37	39.12	36.56	38.37	42.46
		Our Method II	41.74	34.40	37.99	40.14	38.82	45.56
	Llama fine-tuned	TextOnly	46.18$_{1.6}$	44.83$_{1.1}$	46.41$_{1.4}$	38.21$_{1.3}$	43.90	57.46$_{1.7}$
		TextEnty	47.41$_{1.7}$	45.37$_{1.5}$	46.69$_{1.6}$	39.18$_{1.2}$	44.66	58.36$_{1.8}$
		TextRel	46.91$_{1.5}$	46.53$_{1.4}$	47.73$_{1.3}$	40.15$_{1.2}$	45.33	62.17$_{1.4}$
		Our Method II	48.78$_{1.5}$	49.46$_{1.3}$	48.23$_{1.3}$	41.00$_{0.9}$	46.87	63.65$_{1.5}$

Table 7: Mean macro-F1 results (with std) on GCDC and CoheSentia.

Prompt. In the Prompt setting, the data is organized in the Alpaca format (Dubois et al., 2023). Our implementation is built on LlamaFactory (Zheng et al., 2024), a unified framework that incorporates a range of state-of-the-art efficient training methods for large language models (LLMs). In the zero-shot setting, we do not train the models; instead, we directly use LlamaFactory for evaluation. In the fine-tuned setting, we train using LoRA with a rank of 24, a LoRA alpha of 48, a dropout rate of 0.1, a learning rate of 5e-5, and a total of 3 training epochs.

B.3 Baselines

TextOnly. This baseline relies solely on textual content for coherence assessment. In the Fusion setting, we first use a text encoder to generate sentence representations, which are then passed through a sentence-level Transformer for feature extraction and finally fed into a Softmax layer for classification. Notably, no entities or discourse relations are used in this process. In the Prompt setting, we evaluate coherence by inputting only the text into large language models (LLMs). The prompt template used is shown in Figure 6.

TextEnty. This baseline is an ablated version of our approach. In the Fusion setting, we remove discourse relation elements from the flat structure, retaining only sentences and entities. In the Prompt setting, we include only triples connected by entity relations, such as (s_i, entity, s_j), in the prompt.

TextRel. This baseline is another ablated version of our approach. In the Fusion setting, we remove entity elements from the flat structure, retaining only sentences and discourse relations. In the Prompt setting, we include only triples connected by discourse relations, such as (s_i, reason, s_j), in the prompt.

Model			P1	P2	P3	P4	P5	P6	P7	P8	Avg
Fusion	RoBERTa	TextOnly	$74.92_{1.7}$	$70.83_{1.8}$	$74.50_{1.5}$	$75.68_{1.8}$	$76.34_{1.7}$	$72.64_{1.6}$	$72.14_{1.6}$	$71.97_{1.3}$	73.63
		TextEnty	$75.18_{1.8}$	$72.36_{1.5}$	$74.06_{1.4}$	$76.26_{1.2}$	$76.57_{1.7}$	$74.62_{1.6}$	$75.42_{1.6}$	$73.68_{1.7}$	74.77
		TextRel	$75.00_{1.9}$	$72.70_{1.9}$	$75.68_{1.8}$	$74.94_{1.6}$	$76.70_{1.7}$	$72.86_{1.9}$	$73.85_{1.6}$	$73.76_{1.5}$	74.44
		Our Method I	$\mathbf{78.63}_{0.9}$	$\mathbf{75.33}_{1.5}$	$\mathbf{77.98}_{0.6}$	$\mathbf{77.11}_{1.6}$	$77.68_{0.6}$	$\mathbf{77.23}_{1.3}$	$\mathbf{75.90}_{1.9}$	$\mathbf{74.82}_{1.5}$	**76.84**
	Llama	TextOnly	$70.52_{1.7}$	$68.29_{1.3}$	$70.91_{0.8}$	$70.50_{1.6}$	$72.42_{1.4}$	$71.25_{2.1}$	$70.46_{1.3}$	$68.72_{1.7}$	70.38
		TextEnty	$72.39_{1.3}$	$70.66_{1.9}$	$72.71_{1.6}$	$72.13_{1.8}$	$73.50_{1.8}$	$73.53_{1.5}$	$71.29_{1.8}$	$69.37_{1.6}$	72.11
		TextRel	$72.30_{1.5}$	$71.59_{1.3}$	$72.98_{0.6}$	$72.12_{1.8}$	$72.36_{1.8}$	$72.50_{1.8}$	$71.41_{1.5}$	$70.57_{1.4}$	71.98
		Our Method I	$74.30_{1.4}$	$73.97_{2.0}$	$74.48_{1.1}$	$73.76_{1.4}$	$75.48_{2.4}$	$75.96_{1.6}$	$73.82_{1.8}$	$72.54_{2.0}$	74.16
Prompt	Llama zero-shot	TextOnly	45.48	50.80	49.15	47.17	40.96	48.88	41.58	47.17	46.40
		TextEnty	51.48	48.48	51.27	49.16	58.48	52.95	52.26	50.48	50.57
		TextRel	50.37	50.14	51.09	50.64	51.28	51.76	52.56	50.15	51.00
		Our Method II	51.89	50.70	52.73	50.87	51.77	53.06	53.32	51.35	51.96
	Llama fine-tuned	TextOnly	$74.92_{1.7}$	$70.83_{1.8}$	$74.50_{1.5}$	$75.68_{1.8}$	$76.34_{1.7}$	$72.64_{1.6}$	$72.14_{1.6}$	$71.97_{1.3}$	73.63
		TextEnty	$75.18_{1.8}$	$72.36_{1.5}$	$74.06_{1.4}$	$76.26_{1.2}$	$76.57_{1.7}$	$74.62_{1.6}$	$75.42_{1.6}$	$73.68_{1.7}$	74.77
		TextRel	$75.00_{1.9}$	$72.70_{1.9}$	$75.68_{1.8}$	$74.94_{1.6}$	$76.70_{1.7}$	$72.86_{1.9}$	$73.85_{1.6}$	$73.76_{1.5}$	74.44
		Our Method II	$75.69_{2.0}$	$71.71_{1.8}$	$76.21_{1.3}$	$76.11_{1.7}$	$\mathbf{78.71}_{1.7}$	$74.82_{1.5}$	$73.82_{2.0}$	$74.48_{1.7}$	75.19

Table 8: Mean macro-F1 results (with std) on TOEFL dataset.

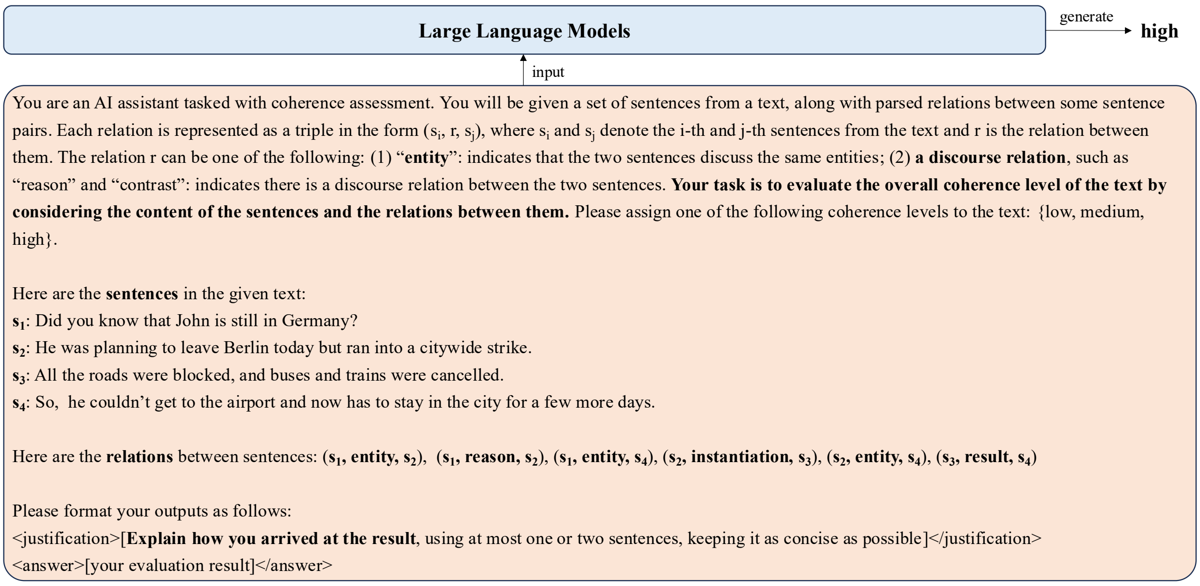

Figure 7: Prompt with explanation.

C Macro-F1 Results

As noted in Section 5, the labels in the GCDC, CoheSentia, and TOEFL corpora are imbalanced. While accuracy is commonly used as the evaluation metric for coherence assessment (Lai and Tetreault, 2018; Jeon and Strube, 2020) and many other NLP tasks (Fu and Frank, 2023, 2024b,a), it does not account for the uneven label distribution (Liu et al., 2019a, 2021). To address this, we also report model performance using Macro-F1, a standard metric for evaluating imbalanced datasets (Opitz and Burst, 2019). Tables 7 and 8 present the results on the GCDC, CoheSentia, and TOEFL datasets. The trends in Macro-F1 scores closely mirror those observed in accuracy: incorporating entities and discourse relations improves performance, and combining both yields the best results.

D Prompt with Explanation

In the case studies presented in Section 5, we prompt LLaMA not only to evaluate the coherence level of a given text but also to provide a brief explanation for its judgment. This is done by modifying the instruction template used with LLaMA. Figure 7 shows the prompt used in these case studies for Our Method II. Similar prompts are used for TextOnly, TextEnty, and TextRel.

E Zero-shot results using LLama-3.3-70B

Coherence assessment involves processing entire documents as input, which are typically quite

Model			GCDC					CoheSentia
			Clinton	Enron	Yahoo	Yelp	Avg	
Prompt	Llama-3.3-70B zero-shot	TextOnly	56.50	51.00	43.50	47.50	49.63	55.07
		TextEnty	57.50	51.50	45.50	52.00	51.63	56.11
		TextRel	59.50	52.50	49.50	52.50	53.50	56.73
		Our Method II	60.00	53.50	52.50	53.00	54.75	57.56

Table 9: Mean accuracy results of **Llama-3.3-70B** on GCDC and CoheSentia in the **zero-shot setting**.

Model			GCDC					CoheSentia
			Clinton	Enron	Yahoo	Yelp	Avg	
Prompt	Llama-3.3-70B zero-shot	TextOnly	41.84	36.30	36.12	35.55	37.45	45.84
		TextEnty	44.61	38.68	40.74	38.68	40.68	48.74
		TextRel	45.69	41.42	42.83	39.74	42.42	48.46
		Our Method II	47.00	40.68	41.69	41.56	42.73	50.62

Table 10: Mean macro-F1 results of **Llama-3.3-70B** on GCDC and CoheSentia in the **zero-shot setting**.

Models			P1	P2	P3	P4	P5	P6	P7	P8	Avg
Prompt	Llama-3.3-70B zero-shot	TextOnly	57.25	58.51	54.58	54.67	57.95	56.46	53.62	54.37	55.93
		TextEnty	60.51	58.26	56.30	58.05	58.25	60.42	60.26	56.80	58.61
		TextRel	61.05	59.35	56.88	58.45	59.83	60.21	61.33	56.51	59.20
		Our Method II	62.56	60.24	59.74	59.91	61.35	62.19	61.80	58.23	60.75

Table 11: Mean accuracy results of **Llama-3.3-70B** on TOEFL dataset in the **zero-shot setting**.

Models			P1	P2	P3	P4	P5	P6	P7	P8	Avg
Prompt	Llama-3.3-70B zero-shot	TextOnly	48.28	52.18	51.06	49.55	48.29	52.45	48.43	51.70	50.24
		TextEnty	51.42	51.69	53.36	52.34	51.72	55.06	54.21	53.62	52.93
		TextRel	52.37	53.42	53.87	53.82	52.55	54.87	56.45	53.88	53.90
		Our Method II	54.01	54.38	55.64	54.28	54.84	56.07	57.35	55.16	55.22

Table 12: Mean macro-F1 results of **Llama-3.3-70B** on TOEFL dataset in the **zero-shot setting**.

lengthy (see Table 6). As a result, training and inference require GPUs with substantial memory capacity. Due to hardware limitations, we employ LLaMA-3.1-8B as the language model for implementing Method II in Section 4. Although we also experimented with the more advanced LLaMA-3.3-70B model, it caused out-of-memory errors during fine-tuning. However, our GPU is capable of running LLaMA-3.3-70B in a zero-shot setting for Method II. Accordingly, we report the zero-shot results (including Accuracy and Macro-F1) using LLaMA-3.3-70B in Tables 9, 10, 11, and 12. As shown, the results are consistent with those obtained using LLaMA-3.1-8B: incorporating entity and discourse relations improves the model's performance in coherence assessment, and jointly modeling both types of information yields the best results.