DECOR: Improving Coherence in L2 English Writing with a Novel Benchmark for Incoherence Detection, Reasoning, and Rewriting

Xuanming Zhang[1], Anthony Diaz[2], Zixun Chen[1], Qingyang Wu[1], Kun Qian[1], Erik Voss[1], Zhou Yu[1]

[1]Columbia University, [2]University of California, Davis

{xz2995, zc2738, zy2461}@columbia.edu, antdiaz@ucdavis.edu

Abstract

Coherence in writing, an aspect that second-language (L2) English learners often struggle with, is crucial in assessing L2 English writing. Existing automated writing evaluation systems primarily use basic surface linguistic features to detect coherence in writing. However, little effort has been made to correct the detected incoherence, which could significantly benefit L2 language learners seeking to improve their writing. To bridge this gap, we introduce *DECOR*, a novel benchmark that includes expert annotations for detecting incoherence in L2 English writing, identifying the underlying reasons, and rewriting the incoherent sentences. To our knowledge, *DECOR* is the first coherence assessment dataset specifically designed for improving L2 English writing, featuring pairs of original incoherent sentences alongside their expert-rewritten counterparts. Additionally, we fine-tuned models to automatically detect and rewrite incoherence in student essays. We find that incorporating specific reasons for incoherence during fine-tuning consistently improves the quality of the rewrites, achieving a result that is favored in both automatic and human evaluations.[1]

1 Introduction

Automatic English writing tools have gained extensive popularity among second-language (L2) learners. These tools serve as a cost-effective supplement to traditional, expensive human tutoring, providing learners with timely and constructive feedback. Much progress in this area includes automatic grammar correction systems (Omelianchuk et al. 2020; Yasunaga et al. 2021; Tarnavskyi et al. 2022; Cao et al. 2023) and tools to improve the vocabulary usage of learners (Johnson et al. 2016; González 2017; Zhang et al. 2024). However, these tools primarily focus on the word and sentence-level issues that affect L2 writing rather than discourse-level issues.

An aspect of L2 writing that could also benefit from automated tools is the overall textual coherence which is a requirement to efficiently convey one's ideas. To improve L2 writing skills, whether it is part of a course assessment or standardized test of English ability, learners are often required to carefully organize their thoughts in response to a predetermined writing prompt. Previous research has identified coherence as a crucial feature to measure when assessing L2 writing proficiency, as it is an aspect that students often struggle with (Schneider and Connor 1990; Bitchener and Basturkmen 2006; Cooley and Lewkowicz 1995; Lorenz 1999). Current automated writing evaluation tools primarily provide learners with scores that indicate the level of coherence in their writing (Naismith et al., 2023). They primarily detect coherence with simple surface linguistic features, such as syntax and parts of speech (McNamara et al., 2010; Crossley et al., 2016). However, merely detecting coherence in writing is insufficient to help L2 English writers enhance their writing. An automated system capable of detecting incoherence in L2 writing, identifying the underlying reasons, and correcting the incoherent sentences would be immensely valuable for both language learners and instructors. However, the absence of a benchmark dataset specifically designed for incoherence detection, reasoning, and rewriting in L2 English essays significantly impedes the development of such systems.

Hence, we introduce *DECOR*, a novel benchmark dataset that can be used to improve coherence in L2 English writing. To construct *DECOR*, we start by creating context-sentence pairs from the TOEFL-11 corpus (Blanchard et al., 2013), following the incremental annotation protocol suggested by Maimon and Tsarfaty (2023). We then design a language-learning-oriented annotation scheme that guides expert annotators to detect incoherence in

[1]Data and code available: https://github.com/BillyZhang24kobe/writing2coherence

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11436–11458
November 12-16, 2024

these pairs, identify specific reasons for incoherence, and rewrite the incoherent sentences. Figure 1 demonstrates the overview of *DECOR* and the three tasks. To our knowledge, *DECOR* is the first benchmark to feature expert annotations for incoherence detection, reasoning, and rewriting, specifically tailored for L2 English writing. The resulting parallel corpus with pairs of original incoherent sentences and their expert-revised versions, provides a valuable resource for evaluating coherence in automated writing evaluation systems.

Moreover, while previous research demonstrated the effectiveness of using GPT-4 to assess writing coherence (Naismith et al., 2023), challenges persist, particularly for users in developing countries (Bubeck et al., 2023; Firdaus et al., 2023). These include limited access to GPT-4, the high cost associated with its usage, and its tendency to produce overly invasive and non-essential revisions (as shown in Figure 1). Consequently, building smaller and more accessible models tailored specifically to our dataset could bring significant benefits. Hence, we develop models to automatically perform the proposed three tasks on *DECOR*. The findings from our experiments indicate that our incoherence detection models deliver performance comparable to GPT-4 in zero-shot and few-shot scenarios, despite being significantly smaller and less costly. We also demonstrate that both automatic and human evaluations affirm that fine-tuning rewriting models with specific reasons for incoherence consistently enhances their ability to produce rewrites that match the quality of those generated by human annotators.

Overall our contributions are three-fold:

- We introduce a novel benchmark *DECOR*, comprising 1,352 context-sentence pairs, which can serve as a valuable resource for assessing coherence in automated writing evaluation systems.
- We produce the first parallel corpus that includes 213 pairs of original incoherent sentences as well as their expert-rewritten counterparts.
- We fine-tuned models using task-specific synthetic data and evaluated them on *DECOR*. These models achieve results comparable to GPT-4 in detecting incoherence and producing rewrites that match the quality of those generated by human experts.

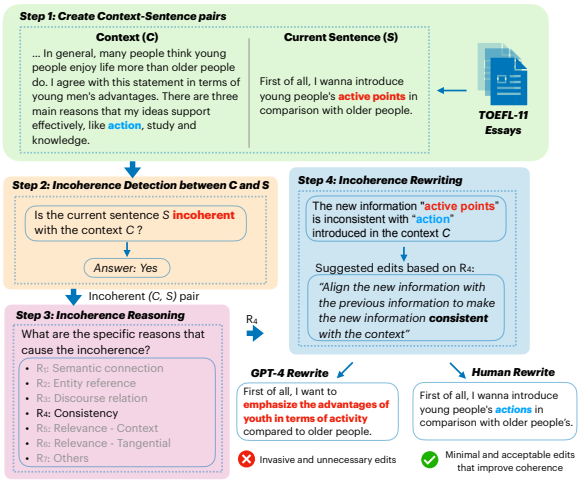


Figure 1: The overview of *DECOR*, containing three tasks: incoherence detection, reasoning, and rewriting. An example human rewrite is generated for the given context-sentence pair. GPT-4 rewrite is unacceptable since it generates more invasive and unnecessary changes.

2 Related Work

2.1 Definitions of coherence in English writing

Earlier efforts at defining coherence in English, such as Halliday and Hasan (1976), focus on explicit cohesive ties (e.g. semantic relations between elements). In particular, Halliday and Hasan (1976) define cohesion as a combination of lexical and grammatical items that facilitate sentences to be understood as connected discourse rather than individual sentences. Moreover, Lautamatti (1978) defined Topical Structure Analysis (TSA) that focuses on different types of progression that are used to create coherence in a text to advance the discourse topic (Knoch, 2007). Additionally, Reinhart (1980) introduced three conditions for a text to be coherent: cohesion, consistency, and relevance, capturing various aspects of the text. In developing our annotation scheme, we referred to these previous efforts and established a useful guideline that is beneficial for annotating incoherence in L2 English writing.

2.2 Assessing coherence in machine-generated texts and human-written texts

Machine-generated texts Following the linguistic definition of coherence established in Reinhart (1980), a more recent work by Maimon and Tsarfaty (2023) incorporated these conditions into a novel benchmark, namely CoheSentia, and proposed a new coherence-annotation protocol that

aligns better with human judgments. Unlike previous work that assigns a single holistic coherence score to each target text (Lai and Tetreault, 2018), CoheSentia provides incremental coherence labeling on a sentence-by-sentence basis, enabling humans to identify the specific reasons for incoherence. In our human annotation process, we follow the CoheSentia protocol to create the context-sentence pairs incrementally. We expand the linguistic fundamentals applied in CoheSentia and devise an annotation scheme that is tailored to incoherence detection and rewriting in L2 English writings.

Human-written texts NLP techniques of Coherence detection for human-written texts primarily identified simple surface feature proxies. McNamara et al. (2010) developed Coh-Metrix that measures cohesion from a wide range of linguistic indexes. Similarly, Crossley et al. (2016) proposed a toolkit for automatic analysis of text cohesion. Recent work by Naismith et al. (2023) investigated the ability of GPT-4 to produce ratings for discourse coherence assessment.

3 *DECOR* Benchmark and Annotation Scheme

In this section, we detail the data creation process for *DECOR* (Section 3.1). We also outline the specific annotation schemes for each proposed task: Incoherence Detection (Section 3.2), Incoherence Reasoning (Section 3.3), and Incoherent Sentence Rewriting (Section 3.4).

3.1 Data Creation

We propose *DECOR*, a benchmark for assessing the writing coherence in L2 English essays. To construct the dataset, we first sampled 100 medium-level essays from the TOEFL-11 dataset (Blanchard et al., 2013). Note that sentences from the TOEFL-11 dataset often have basic grammar mistakes like spelling errors or missing be-verbs, which can make the intended meaning unclear. Therefore, we used a grammar error correction model from Zhang et al. (2024) to fix these mistakes without changing the overall meaning of the sentence. Then, we incrementally constructed context-sentence pairs (C, S) for each essay, following the protocol suggested by Maimon and Tsarfaty (2023). In these pairs, sentence S is the current sentence to be assessed, and context C includes all preceding sentences in the essay up to and including the sentence immediately before S. Overall, we constructed $1,352$ (C, S) pairs from the 100 essays. The general statistics of *DECOR* are shown in Table 1. More detailed statistics, such as the number of sentences and words per essay, are shown in Figure 5 in the Appendix. Next, for each context-sentence pair (C, S), we ask our human annotators to complete three tasks according to our annotation schemes: incoherence detection, reasoning, and rewriting. These three tasks are the main features of *DECOR*. We discuss these features and their specific annotation schemes below.

Items	Count
# of essays	100
# of words	26,376
# of context-sentence pairs	1,352
# of coherent sentences	906
# of incoherent sentences	446
# of human rewrites	213

Table 1: Overall statistics of *DECOR*.

3.2 Incoherence Detection Annotation Scheme

Inspired by the linguistic fundamentals of coherence (i.e. cohesion, consistency, and relevance) defined in Reinhart (1980), we expanded these fundamentals with reference to previous work in order to apply the task of incoherence detection to L2 English writing. We describe five specific criteria for detecting incoherence in each context-sentence pair below.

Semantic connection serves as the criterion that is based on the expanded categories of discourse progression for TSA proposed in Lautamatti (1978), where a sentence's semantic connection with the context of discourse is defined by its appropriate use of the sequential progression of topics from sentence to sentence that contributes to local coherence (Reinhart 1980; Knoch 2007). **Entity reference** refers to the requirement for writers to establish a link between the topics of the current sentence and the context of the discourse and is related to cohesion. Accurate anaphoric pronominal use is a key component of this criterion (Knoch, 2007). For instance, in the passage *Learning about ideas and concepts is essential for all students. For example, they help students to apply their knowledge in new ways.*, the pronoun *they* in the second sentence agrees in person and number with the referent *ideas and concepts* in the first sentence.

Label Codes	Descriptions	Examples
R1: Semantic connection	The sentence S does not connect semantically with the context C.	C: If students study ideas and concepts, they can explore new areas of research. S: We need to make effort to apply our knowledge S': ***They*** need to make effort to apply ***their*** knowledge.
R2: Entity reference	The current sentence S discusses an entity that has not been introduced in C yet, or sentence S discusses an entity that is ambiguous in C.	C: Some people enjoy tours. S: Guides provide a lot of value for tourists. S':***Traveling in tour groups*** provides a lot of value for ***them***.
R3: Discourse relation	The relation between sentence S and previous ones in C doesn't make sense due to a missing discourse marker.	C Advertisements are not good for consumers. S: They only show the good features of a product. S': ***For example***, they only show the good features of a product.
R4: Consistency	The current sentence S contradicts or is inconsistent with previously presented information.	C: Because gas is getting more expensive, less people will drive in the future. S: Scientists are finding ways to make gas cheaper for drivers. S': Scientists are ***researching alternative sources of energy***.
R5: Contextual relevance	The current sentence S introduces information that is completely irrelevant to the context.	C: To become successful, people need to take risks. S: I think fear controls our decision making process. S': ***Risks are important for people to learn what works and what doesn't work***.
R6: Tangential relevance	The current sentence S introduces information that is tangential or unnecessary for the development of the context.	C: Young people tend to not help the people of their community. S: When I was younger I used to volunteer at a retirement home. S':***As a result, there may be a lack of volunteers an places like retirement homes***.
R7: Others	Other reasons that are not listed above. For example, the comment (rheme/focus) of the sentence does not agree with the topic of the sentence.	S: My pet fish is flying in the sky. S': My pet fish is ***swimming*** in ***its tank***.

Table 2: Label codes for the specific reasons for incoherence during annotation. The rewrites S' are provided for each incoherent (C, S) pair. The erroneous parts in S are marked in red, and the corrections are marked bold in S'.

Discourse relation is concerned with how the sentence is related to the overall context through the use of explicit cohesive ties that refer to the semantic relations between an element in a text and some other element that is crucial to the interpretation of it (Halliday and Hasan, 1976). **Consistency** is associated with the logical requirements for a sentence to align with the preceding sentences in the context (Maimon and Tsarfaty, 2023). **Relevance** dictates a sentence must be related to previous sentences in the discourse and the underlying discourse topic of the global context (Maimon and Tsarfaty, 2023).

If the given context-sentence pair violates any of the aforementioned criteria, it is considered incoherent, necessitating the subsequent step (described in Section 3.3) to identify the specific reasons causing sentence S to be incoherent to context C; otherwise, the sentence is labeled as coherent. The detailed annotation guidelines for this task are demonstrated in Appendix A. Note that annotators are instructed to evaluate the entire context when determining if a sentence is incoherent. If the context pertains to a fictional setting, such as a dream about fiction, these instances will not be considered incoherent.

3.3 Incoherence Reasoning Annotation Scheme

In addition to detecting incoherence, annotators are tasked with identifying the specific reasons for incoherence in the context-sentence pairs that are labeled as such. Drawing on the linguistic principles of coherence outlined in Reinhart (1980), three primary factors contribute to incoherence: *Cohesion*, *Consistency*, and *Relevance*. Given that *Cohesion* pertains to the linear sequencing and connections of sentences, we specifically designated three label codes for annotations within this category: semantic connection, entity reference, and discourse relation. For *Consistency*, we use a single code: consistency. Regarding *Relevance*, we have devised two codes: contextual relevance and tangential relevance. Other possible reasons that are not listed above are referred to as others. Detailed descriptions and examples of each label code are illustrated in Table 2.

3.4 Incoherent Sentence Rewriting Annotation Scheme

After selecting all applicable reasons, sentence S is rewritten by the annotators to convert it to be coherent with context C. Concretely, annotators are asked to make the least invasive changes necessary to improve the coherence based on the identified reasons. For example, if *Discourse relation* is selected as the reason, annotators are instructed to *add or change a discourse marker that ties sentence S with context C*. The complete list of suggested edits is described in Appendix A.2. Considering the challenges of providing all possible edits to sentence S during the annotation process, we instructed our annotators to provide only one possible edit that addresses at least one selected reason from the previous step. We leave the exploration of multiple edits for future work.

4 Data annotation process and statistics

Considering the need for substantial experience in English essay grading, we recruited two annotators with extensive teaching experience in English

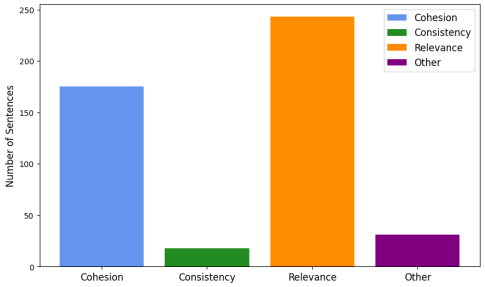
and advanced degrees in Applied Linguistics, specializing in English language education. Before annotating *DECOR*, we conducted a tutorial session to train the two annotators and familiarize them with our annotation scheme. Subsequently, in accordance with our specified scheme, we tasked them with annotating five sample essays, which comprised 72 sentence-context pairs.

We calculated the inter-annotator agreement for these pairs using Cohen's Kappa (Cohen, 1960). The two annotators achieved a κ value of 0.83 for Incoherence Detection, indicating an almost perfect agreement. For Incoherence Reasoning, they reached an average $\kappa = 0.90$ across all reason types, also reflecting almost perfect agreement. The specific agreement scores for each reason type and more justifications for the annotation process are presented in Appendix B. As for Incoherent Sentence Rewriting, the leading authors validated whether the new sentences are acceptable. In particular, a new sentence S' is acceptable if it preserves the semantic meaning of the original sentence S and is coherent with the given context C. Overall, the rewrites by the two annotators were deemed acceptable at rates of 88% and 89%, respectively.

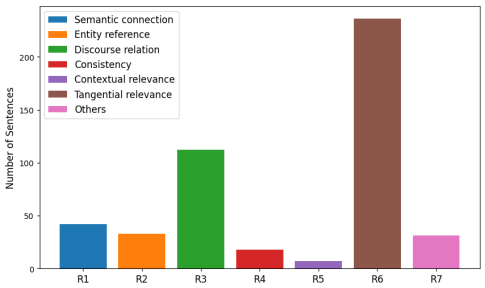
Subsequently, the two annotators worked independently on the test set, with each annotating around 700 (C, S) pairs that are constructed from Section 3. Overall, among all $1,352$ (C, S) pairs, 906 sentences are coherent with their corresponding contexts, whereas 446 sentences are labeled as incoherent. We present the number of words per rewrite in Figure 4. Note that we do not consider rewrites marked as *DELETE*, resulting in 213 rewrites that contain more than one word. In addition, we presented the distribution of the annotated reasons for incoherence in Figure 2. Our analysis shows that the medium-level essays, randomly sampled from the TOEFL-11 corpus, generally maintain consistency and rarely contradict the context. Moreover, we also find that the primary sources of incoherence in these essays are related to *Relevance* and *Cohesion*, with issues of tangential relevance and weak discourse relations being the most prevalent.

5 Incoherence Detection, Reasoning and Rewriting

We propose *DECOR* to benchmark the model's ability in incoherence detection, reasoning, and rewriting for English essays written by L2 language learners. In this section, we will outline each of the three tasks and describe their specific task formulations, evaluation metrics, data, baselines, and results and analysis.



(a) Distribution of reasons for incoherence clustered into groups.



(b) Distribution of specific reasons for incoherence.

Figure 2: Distribution of specific reasons for incoherence, and those clustered into groups.

5.1 Incoherence Detection

5.1.1 Task formulation

In this task, the model will assess the given context-sentence pairs that are extracted from essays written by L2 learners, determining whether the sentence S maintains coherence with the context C. This task is specifically designed to evaluate the effectiveness of systems in capturing coherence within learner-written texts.

5.1.2 Evaluation metrics

Given the class imbalance in our test set, where 906 instances are labeled as coherent and 446 as incoherent, we opt to use the weighted F1 score as a metric to assess the performance of different models. This approach ensures a fair evaluation by accounting for the disproportionate distribution of classes.

5.1.3 Data

Given the absence of a dedicated incoherence detection corpus for language learners suitable for model training purposes, we followed the approach

Models		Incoherence Detection (%)	Incoherence Reasoning (%)			
			Cohesion	*Consistency*	*Relevance*	*Others*
BERT-base	D_C	63.04	48.17	93.76	28.47	-
	D_T	66.43	44.38	75.41	46.37	80.36
DeBERTa-base	D_C	62.21	47.93	**93.88**	29.45	-
	D_T	68.54	48.36	77.17	45.14	74.20
Llama2-7B	D_C	59.52	43.93	93.65	28.87	-
	D_T	66.08	46.63	83.55	47.20	87.78
GPT-4	*zero*	66.56	**51.03**	93.02	56.60	**87.93**
	16	**69.33**	48.71	90.10	**65.54**	85.64

Table 3: Evaluation of models on *DECOR* using weighted F1 scores in percentages (%) for Incoherence Detection and Incoherence Reasoning tasks. For each task, the task-specific synthetic training data is denoted as D_T, whereas the out-of-domain training data is denoted as D_C. We also conducted zero-shot (*zero*) and in-context learning (16-shot) with GPT-4. Since *Others* is not specified in D_C, we exclude it for evaluation.

recommended by Zhang et al. (2024) and synthesized task-specific incoherence detection data using GPT-4 (OpenAI, 2023).[2] The prompt we used for GPT-4 is shown in Appendix E.1. To start with, we randomly sampled 800 medium-level essays from the TOEFL-11 dataset and generated 11,267 context-sentence pairs. We then used GPT-4 to analyze these pairs for incoherence, producing a label for each. In this process, 6,422 sentences were identified as coherent, while 4,845 were labeled as incoherent. For the training process, we allocated 90% of this synthetic data for training purposes, denoted as D_T, and reserved the remaining 10% for validation. Moreover, we also utilized out-of-distribution (OOD) training data proposed in Maimon and Tsarfaty (2023), denoted as D_C. To our knowledge, D_C is the only dataset featuring human annotations for coherence detection and incoherence reasoning in machine-generated texts. We incorporated D_C as OOD training data to assess whether models trained exclusively on it could achieve strong performance in *DECOR*. This experiment aims to determine whether using OOD data can improve the detection of incoherence in texts authored by L2 speakers.

5.1.4 Baselines

We conducted experiments with classification-based models that consist of encoder-only architectures equipped with a classification head. Specifically, we tested models such as BERT (Devlin et al., 2018) and DeBERTa (He et al., 2021) with their base and large variants. Each model generates predictions with two labels—*yes* or *no*—to determine if the sentence S is coherent with the context C. The input to the model's encoder is structured in the format "C <SEP> S," facilitating the assessment of coherence between the given context and sentence.

In light of the burgeoning field of powerful instruction-following models (Ouyang et al., 2022; Bai et al., 2022; OpenAI, 2023), we also explored two generation-based large language models: Llama 2 (Touvron et al., 2023) and GPT-4. For Llama 2, we fine-tuned its 7B variant using our synthetic dataset D_T for this task. With GPT-4, we tested in both zero-shot and 16-shot settings. Details of the prompts used in the GPT-4 experiments are provided in Appendix G.

5.1.5 Results and analysis

The results for the task of incoherence detection are demonstrated in Table 3. As observed, training with our task-specific synthetic dataset D_T yielded superior results compared to using the OOD dataset D_C. This improvement is attributed to the fact that D_C consists solely of machine-generated texts, which introduces a significant distribution shift. Additionally, while GPT-4 with 16-shot examples surpassed all other models, smaller models trained on our synthetic data D_T, such as BERT-base and Llama-2-7B, achieved performance comparable to GPT-4 in a zero-shot setting. Moreover, DeBERTa-base matched GPT-4's performance in the 16-shot setting and even exceeded it in the zero-shot scenario. We also experimented with combining both D_C and D_T during training; however, this did not lead to improved results. Details of the experiment are provided in Appendix F.

[2]Throughout this paper, we employ `GPT-4o` as the default model unless otherwise specified.

5.2 Incoherence Reasoning

5.2.1 Task formulation

The incoherence reasoning task aims to develop models capable of identifying the specific causes of incoherence in context-sentence pairs labeled as such. Due to the sparse distribution of incoherence reason types depicted in Figure 2b, we focus on the four high-level causes previously introduced: *Cohesion*, *Consistency*, *Relevance*, and *Others*. For each of these four causes of incoherence, we developed specialized models capable of determining whether the incoherence stems from a specific cause. This approach divides the overall incoherence reasoning task into four distinct sub-tasks, each targeting a different cause.

5.2.2 Evaluation metrics

In Figure 2a, *DECOR* exhibits class imbalance across the four reason types of incoherence. Hence, we report weighted F1 scores for each of the four sub-tasks to account for this imbalance.

5.2.3 Data

We adopted a similar approach as described in Section 5.1.3 to synthesize the training data for all four sub-tasks. Specifically, we prompted GPT-4 to identify all potential reasons for each instance of incoherence detected from Section 5.1.3, based on the seven predefined causes outlined in Table 2. The prompts we used for data synthesis are demonstrated in Appendix E.1. Furthermore, we post-processed the resulting data to create four distinct datasets, each serving as the training data for detecting *Cohesion*, *Consistency*, *Relevance*, and *Others*. For instance, in creating the training set for detecting *Cohesion* as the cause, an instance is labeled "Yes" if GPT-4 identifies R1, R2, or R3 as the cause of incoherence for that instance; otherwise, the label is "No", indicating that the incoherence is caused by other factors. Similar to 5.1.3, the synthetic datasets are denoted as D_T. The details for the post-processing and statistics of the resulting data for each sub-task are described in Appendix E.2.

5.2.4 Baselines

We adopted the same set of baseline models that are tested in the incoherence detection task: classification-based models (i.e. BERT and DeBERTa), and generation-based models (i.e. Llama 2 and GPT-4). Similarly, for each sub-task of the incoherence reasoning, each model predicts with two labels (i.e. yes or no) to determine if the sentence S is incoherent with the context C due to a specific cause. We fine-tuned BERT, DeBERTa, and Llama2-7B models on the task-specific synthetic data D_T for each sub-task as well as the out-of-distribution data D_C. We also prompted GPT-4 under both zero-shot and 16-shot settings. The prompts for GPT-4 experiments are shown in Appendix G.

5.2.5 Results and analysis

The results for incoherence reasoning in terms of the four sub-tasks are demonstrated in Table 3. It was observed that training DeBERTa-base and Llama2-7B models with D_T resulted in enhanced performance for *Cohesion* and *Relevance* when compared to training with D_C. For *Cohesion*, DeBERTa-base outperforms the Llama2-7B model and is close to the performance of GPT-4. In comparison, for the *Consistency* task, all of our models demonstrate markedly enhanced performance when trained with D_C rather than D_T. This improvement is likely attributed to the imbalanced training data distribution in D_C, which more closely mirrors the *Consistency* class distribution in *DECOR*. For the task of *Others*, we have omitted D_C from the table because the category *Others* is not included in D_C. Our Llama2-7B model, fine-tuned with D_T, achieved results comparable to GPT-4 in both zero-shot and 16-shot settings. We further explored the effects of combining D_T and D_C as training data to fine-tune our models for tasks excluding *Others*. The results varied across different tasks and are presented in Table 8 in Appendix F.

5.3 Incoherence Rewriting

5.3.1 Task formulation

The incoherence rewriting task is designed to assess the model's capability to edit a given incoherent sentence S to a revised sentence S' that restores the coherence with context C, based on the identified reasons R for incoherence. Specifically, we prefer edits that not only enhance the coherence of the original sentence but also minimize alterations, ensuring the changes are as unobtrusive as possible.

5.3.2 Evaluation metrics

We measured the systems' performance on the task of incoherence rewriting with the acceptance rate. This metric was determined by calculating the proportion of revised sentences S' that both achieve co-

Model	Training Condition	Acceptance Rate (%)	Win Rate (%)
Llama2-7B	w/ reason	75.59	69.16
	w/o reason	74.65	69.01
Llama3-8B-Instruct	w/ reason	**77.46**	**72.30**
	w/o reason	75.12	71.83

Table 4: Automatic evaluation of models for the incoherence rewriting task. The win rate is calculated by adopting GPT-4 as a judge to compare the system-generated rewrites against human-written references.

herence with context C and maintain minimally invasive edits, out of all evaluated incoherent context-sentence pairs. We specifically employed GPT-4 with 16-shot examples (with the best performance in the incoherence detection task) to determine if the rewrites S' are acceptable. Additionally, in line with the recent practices of evaluating instruction-following LLMs (Zhou et al., 2024; Dubois et al., 2024), we asked GPT-4 to rank a pair of generated rewrites (one from the human-written reference, the other from the tested models) to decide which one is more coherent to the context C. For each tested model, we collect its win rate against the human reference. Note that we randomly shuffle the ordering of the pair-wise outputs to avoid position biases. The prompt we adopted for GPT-4 judging is shown in Appendix H.

5.3.3 Data

Given the reasons generated from the incoherence reasoning task, we prompted GPT-4 to generate the rewrites based on the identified reasons for incoherence. These rewrites are used as the training data for the incoherence rewriting task. The prompt we used for the rewrite synthesis and relevant statistics are shown in Appendix E.1. For automatic evaluation, we used all 213 rewrites generated by our annotators, and we randomly selected a sample of 100 for human evaluation.

5.3.4 Baselines

We conducted experiments with two advanced open-sourced generative LLMs, Llama 2 (Touvron et al., 2023) and Llama3 (AI@Meta, 2024), for the incoherence rewriting task. Specifically, we fine-tuned Llama2-7B and Llama3-8B-Instruct using our synthetic rewriting dataset under two experimental conditions: training with reasons for incoherence and without reasons.

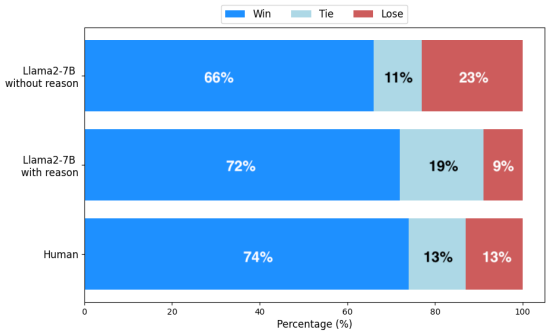


Figure 3: Human expert as a judge evaluation results with GPT-4 rewrites as the baseline. We sample 100 examples and ask our human expert for each pair of comparisons. A higher win rate and a lower loss rate indicate superior quality.

5.3.5 Results and analysis

Automatic Evaluation The automatic evaluation results for incoherence rewriting are shown in Table 4. As observed, fine-tuning both the Llama2-7B and Llama3-8B-Instruct models with reasons for incoherence consistently results in better performance compared to their counterparts trained without such reasons, achieving higher scores in both acceptance rate and win rate. Table 9 demonstrates the qualitative comparisons among example rewrites produced by our fine-tuned models.

Human Evaluation Moreover, we conducted a human evaluation where we asked our human expert to judge and compare system-generated rewrites with those produced by GPT-4.[3] The detailed information for the human evaluation process is shown in Appendix D. Additionally, the human evaluator was also tasked with a pairwise comparison between human-written references and the same set of GPT-4 rewrites. The results are shown in Figure 3. As expected, our human judges predominantly preferred rewrites produced by human experts over those generated by GPT-4, with the highest win rate reaching 74%. Consistent with the results in Table 4, fine-tuning Llama 2 with reasons for incoherence resulted in a higher win rate and a significantly lower loss rate compared to fine-tuning without reasons. A chi-square test indicates a significant difference between these two conditions (with p-value < 0.01). This supports our hypothesis that rewriting incoherent sentences with an understanding of their underlying causes produces higher-quality rewrites.

[3]To avoid biases, instead of the same annotators, we asked one of our leading authors to conduct the human evaluation.

6 Conclusion and Future Work

We propose a novel benchmark *DECOR* aiming to assess and improve coherence in L2 English writing. Specifically, *DECOR* contains three tasks: incoherence detection, reasoning, and rewriting. Our annotation scheme allows us to produce a corpus comprising $1,352$ context-sentence pairs with coherence labels, as well as the first parallel corpus featuring 213 pairs of original incoherent sentences and their expert-rewritten counterparts. Additionally, we fine-tuned various models with task-specific synthetic data, achieving results comparable to GPT-4 in coherence detection and generating rewrites favored by both automatic and human evaluations. In future work, we plan to enhance *DECOR* by expanding its size and quality, ensuring more balanced reasoning types and multiple edits for each incoherent context-sentence pair. This enhancement will create a more comprehensive evaluation set for incoherence detection and correction, specifically tailored to L2 writing.

7 Limitations

While our benchmark may contribute to building the systems that can improve the coherence in L2 English writing, there were a number of limitations to our study.

First, the distribution of incoherence reason types is unbalanced, with the *Consistency* category containing the fewest annotations among the four high-level reason types. This is due to the fact that medium-level essays from the TOEFL-11 corpus, the source of all context-sentence pairs, generally maintain consistency and seldom contradict the context. We leave our future work to diversify and balance the reason types in *DECOR*, potentially by including low-level essays written by English L2 learners.

Additionally, the texts sampled from the TOEFL-11 corpus for synthesizing our training data were limited by the specific writing prompts they addressed. This limitation may hinder the system's ability to detect coherence in learner-produced writing that responds to out-of-domain prompts not included in the TOEFL-11 corpus. Future extensions of our work include incorporating other L2 English writing corpora.

Finally, regarding the general design of our annotation scheme for coherence detection, we considered all sentences in the context up until the target sentence. However, as we found during our annotation tutorial session, sometimes issues of coherence occur due to the structuring of information that is contained in sentences that come later in the text. Future work might focus on these specific types of coherence breaks and their prevalence in L2 writing.

8 Ethics Statement

Reproducibility In this work, we utilized GPT-4 to synthesize our task-specific training data for coherence detection, reasoning, and rewriting. We also used it during the evaluation. To facilitate the reproducibility of our data synthesis process and evaluation results, we included all relevant prompts that were used in our paper. In addition, all the other models used in this research, are publicly available in peer-reviewed articles and referenced in this paper. All datasets, including our synthetic fine-tuning dataset and the annotated test set, are released.

Biases We did not explicitly handle any bias that exists in the pre-trained language models we experimented with in this paper.

Human Annotators Both annotators were specifically recruited from the linguistics department, and they are both associate professors with extensive experience in teaching English as a foreign language and have advanced degrees in Applied Linguistics. They were paid at a rate of $12 per hour. To protect privacy and anonymity, contributors' personal and demographic information was not collected.

9 Acknowledgement

We would like to thank Yanda Chen, Ryan Shea, and people from the Columbia NLP group for their valuable discussions on the paper. We also thank all reviewers for their constructive feedback and suggestions, which significantly improve our work. In addition, we extend our gratitude to our expert annotators for their time and contributions to the completeness of the benchmark.

References

AI@Meta. 2024. Llama 3 model card.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with

reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

John Bitchener and Helen Basturkmen. 2006. Perceptions of the difficulties of postgraduate l2 thesis students writing the discussion section. *Journal of English for Academic Purposes*, 5(1):4–18.

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.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Hannan Cao, Wenmian Yang, and Hwee Tou Ng. 2023. Mitigating exposure bias in grammatical error correction with data augmentation and reweighting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2123–2135.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Linda Cooley and Jo Lewkowicz. 1995. The writing needs of postgraduate students at the university of hong kong: A project report. *Hong Kong papers in Linguistics and language teaching*, 18:121–123.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48:1227–1237.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.

Muhammad Fatihul Firdaus, Joseph Nugraha Wibawa, and Fajri Fathur Rahman. 2023. Utilization of gpt-4 to improve education quality through personalized learning for generation z in indonesia. *IT for Society*, 8(1).

Melanie C González. 2017. The contribution of lexical diversity to college-level writing. *TESOL Journal*, 8(4):899–919.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 1976. *Cohesion in english*. Routledge.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Mark D Johnson, Anthony Acevedo, and Leonardo Mercado. 2016. Vocabulary knowledge and vocabulary use in second language writing. *TESOL Journal*, 7(3):700–715.

Ute Knoch. 2007. 'little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing writing*, 12(2):108–128.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. *arXiv preprint arXiv:1805.04993*.

Liisa Lautamatti. 1978. Observations on the development of the topic in simplified discourse. *AFinLAn vuosikirja*, pages 71–104.

Gunter Lorenz. 1999. Learning to cohere: Causal links in native vs. non-native argumentative writing. *Pragmatics and Beyond New Series*, pages 55–76.

Aviya Maimon and Reut Tsarfaty. 2023. Cohesentia: A novel benchmark of incremental versus holistic assessment of coherence in generated texts. *arXiv preprint arXiv:2310.16329*.

Danielle S McNamara, Max M Louwerse, Philip M McCarthy, and Arthur C Graesser. 2010. Coh-metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292–330.

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.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. Gector–grammatical error correction: tag, not rewrite. *arXiv preprint arXiv:2005.12592*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Melanie Schneider and Ulla Connor. 1990. Analyzing topical structure in esl essays: Not all topics are equal. *Studies in second language acquisition*, 12(4):411–427.

Maksym Tarnavskyi, Artem Chernodub, and Kostiantyn Omelianchuk. 2022. Ensembling and knowledge distilling of large sequence taggers for grammatical error correction. *arXiv preprint arXiv:2203.13064*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2021. Lm-critic: Language models for unsupervised grammatical error correction. *arXiv preprint arXiv:2109.06822*.

Xuanming Zhang, Zixun Chen, and Zhou Yu. 2024. Prolex: A benchmark for language proficiency-oriented lexical substitution. *arXiv preprint arXiv:2401.11356*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A Detailed annotation scheme for *DECOR*

A.1 Incoherence detection and reasoning

In the coherence detection process, coherent (C, S) pairs are marked with a 1, while incoherent ones are marked with a -1. For cases unrelated to writing coherence (e.g., sentence parsing errors), a 0 is assigned and they will be excluded from the resulting dataset.

To complete this task, annotators were instructed that for each sentence there is a topic T, and a context C, which comprises all preceding sentences up to and immediately before sentence S in the essay and for all incoherent sentences to provide all possible reasons (R1-R7) for the break in coherence. They were also instructed to determine if each sentence S is coherent with context C based on the provided instructions. Lastly, for each incoherent sentence S, annotators were asked to revise S to improve its coherence, taking into account the types of edits suggested for each identified reason. Below is the complete list of reasons that were provided to the annotators.

- (1) The sentence S is coherent with the context C as:
  - The sentence S semantically connects to the context C, (i.e. with proper use of reference words, repeated words/ ideas, and substitution), and
  - All entities discussed in the new sentence S have been introduced in C, and
  - The new sentence S demonstrates reasonable discourse relation with previous ones, and
  - The new sentence S contains a meaning consistent with previously presented data in C and
  - The new sentence S contains a meaning relevant to previously presented data in C
- (-1) The sentence S is not coherent with C as:
  - R1: (Semantic connection) The sentence S does not connect semantically with the context C;
  - R2: (Entity reference) The new sentence S discusses an entity that has not been introduced in C yet, or the new sentence S discusses an entity that is ambiguous in C or
  - R3: (Discourse relation) The relation between sentence S and previous ones in C doesn't make sense due to a missing discourse marker.
  - R4: (Consistency) The new sentence S contradicts or is inconsistent with previously presented information, or
  - R5: (Contextual relevance) The new sentence S introduces information that is completely irrelevant to the context
  - R6: (Tangential relevance) The new sentence S introduces information that is either tangential or slightly irrelevant to the context.
  - R7: (Others) Other reasons that are not listed above. For example, the comment (rheme/focus) of the sentence does not agree with the topic of the sentence.
- (0) Other cases that have nothing to do with writing coherence

For incoherent reasons, annotators were asked to mark "1" in the corresponding reason column of the annotation document and leave the others empty. For example, if sentence S is incoherent to context C due to reason 2 (Entity reference) and reason 3 (Discourse relation), mark "1" in both R2 and R3 columns, and leave the others empty.

A.2 Types of edits for incoherent sentence rewriting

Given an incoherent sentence-context pair (C, S), annotators are instructed to make the least invasive changes to rewrite sentence S. The suggested edits are described as follows:

- *Semantic connection*: add reference words or repeated words/ideas or substitution that can semantically connect sentence S to context C.
- *Entity reference*: link the newly introduced entity or ambiguous entity in sentence S to context C.
- *Discourse relation*: add or change a discourse marker that ties sentence S with context C.
- *Consistency*: align the newly introduced information in sentence S with previously introduced information in context C so that the new information does not contradict the context.

- *Contextual relevance*: modify sentence S so that it is relevant to the context established by the writer.
- *Tangential relevance*: delete the sentence and edit with "DELETE".
- *Others*: rewrite the sentence so that the comment of the sentence agrees with the topic of the sentence.

Note that we recommend "DELETE" if sentence S is tangential, as its presence following context C is unnecessary.

B More details about inter-annotator agreement

B.1 Inter-annotator agreement scores across different reason types

The specific inter-annotator agreement scores for both incoherence detection and reasoning tasks are shown in Table 5. Overall, our annotators achieved very high agreement on both tasks.

Group	Cohen's κ
Coherence	0.83
Cohesion	0.80
Consistency	1.00
Relevance	0.86
Others	1.00

(a) Inter-annotator agreement on incoherence detection and reasons clustered into groups.

Reasons	Cohen's κ
R1	0.84
R2	0.74
R3	0.88
R4	1.00
R5	1.00
R6	0.86
R7	1.00

(b) Inter-annotator agreement on specific reasons.

Table 5: Inter-annotator agreement scores for annotations.

B.2 The achievability of high inter-annotator agreement on *DECOR*

We achieve a high inter-rater agreement for our dataset through our meticulously structured and clearly defined annotation process and scheme. Specifically, we recruited two expert annotators who are both professors with extensive experience in teaching English as a foreign language and have advanced degrees in Applied Linguistics. Before starting, we delivered a comprehensive tutorial and thoroughly reviewed our detailed annotation guidelines with the annotators. To familiarize them with our annotation scheme, we selected eight medium-level essays from TOEFL-11, generating 109 context-sentence pairs. The tutorial was structured into three sessions to ensure thorough coverage of all instances. During these sessions, annotators collaboratively labeled the samples. We also refined the annotation scheme as necessary. Additionally, we incorporated examples of errors encountered during these tutorial sessions into the annotation scheme. Subsequently, to quantitatively assess the inter-annotator agreement, we sampled an additional five medium-level essays, distinct from those used in the tutorial sessions, resulting in 72 context-sentence pairs.

Among the 72 samples, the first annotator labeled 35 context-sentence pairs as incoherent, and 30 as coherent. The second annotator labeled 39 pairs as incoherent, and 28 as coherent. Note that the rest were labeled as "uncertain". Therefore, the resulting annotations for the 72 samples were relatively balanced. Given the high inter-annotator agreement achieved by the two annotators, with Cohen's Kappa scores of 0.83 for incoherence detection and 0.90 for incoherence reasoning, we subsequently assigned them to independently annotate the entire test set.

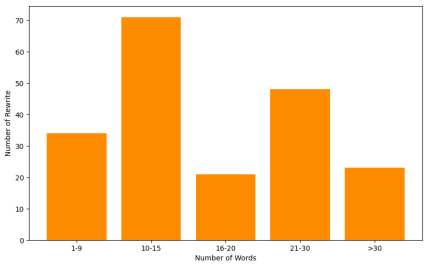
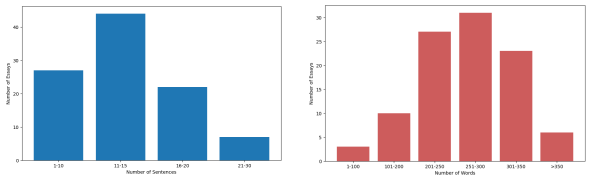


Figure 4: The number of words per rewrite.

C Additional statistics of *DECOR*

We show the overall distribution of rewrite lengths measured by the number of words in Figure 4. We also illustrate the distribution of essays measured by the number of sentences and words in Figure 5.



(a) The number of sentences per essay.

(b) The number of words per essay.

Figure 5: Distribution of essays by number of sentences and number of words.

D Human evaluation details

During human evaluation, our human expert was asked to compare system-generated rewrites with those produced by GPT-4. Specifically, we conducted three sets of pair-wise comparisons: 1) human-generated rewrites VS GPT-4; 2) rewrites generated by Llama2-7B trained with reasons VS GPT-4; and 3) rewrites generated by Llama2-7B trained without reasons VS GPT-4. For each set of pair-wise comparisons, for each pair of outputs (e.g. system 1 output VS system 2 output) the human expert was asked "Between system 1 and system 2 outputs, select the output that provides better rewrite based on the reason for incoherence; otherwise, choose "Tie" to indicate equal quality".

After human evaluation, we interviewed the human expert and asked for his feedback on the system outputs. In general, the evaluator was surprised that our fine-tuned system can generate rewrites of decent quality. Among the three sets of comparisons, he was not able to tell which one was generated by humans. This indicates that our fine-tuned smaller language models are already capable of correcting the incoherence well based on the reasons behind it.

E Synthesizing training data with GPT-4

E.1 Dataset Synthesis

A large portion of the training set of *DECOR* is synthesized by GPT-4 based on human-annotated examples in order to increase generalizability and variety. Table 6 shows the prompt we used.

E.2 Post-Processing

As described in Section 5.2.3, we prompted GPT-4 to identify all potential reasons for each incoherent context-sentence pair. To obtain the training data for detecting *Consistency* as the cause, an instance is labeled as "Yes" if GPT-4 identifies R4 as the cause of incoherence for that instance; otherwise, the label is "No". We conducted similar post-processing steps to create the training data for *Relevance* and *Others* tasks. Given that the initial data after post-processing is extremely unbalanced for each sub-task. We downsampled instances of the majority class to achieve a more balanced training dataset. The statistics of the resulting data for each sub-task are shown in Table 7.

F Details of Experiments

F.1 Classification-based models

For training the BERT and DeBERTa models, we established our pipeline based on the platform developed by (Maimon and Tsarfaty, 2023). Specifically, for the incoherence detection and reasoning tasks, we train these two models on both the CoheSentia dataset D_C and the synthetic training data D_T, as well as a combination of the two, $D_C + D_T$. For validation purposes, we utilized the existing validation dataset from CoheSentia. Additionally, we allocated 10% of the synthetic dataset for evaluating models trained with D_T. Note that since *DECOR* and CoheSentia has different definitions for the Others category, it is not possible to evaluate a model trained with D_C for this category on *DECOR*, nor does it make sense to combine the datasets for this category. All models are trained for 10 epochs on their respective dataset with a learning rate of 2×10^{-5} and batch size of 8 on a single NVIDIA A100-80G GPU. Based on the results from the validation set, we evaluate the best checkpoint on *DECOR* for each task.

F.2 Generation-based models

For the task of incoherence detection and reasoning, we fine-tuned Llama2-7B under three experiment settings (i.e. D_T, D_C, and $D_C + D_T$). For the task of incoherence rewriting, besides Llama2-7B, we additionally fine-tuned Llama3-8B-Instruct on our synthetic training data generated by GPT-4. Specifically, we referred to the platform developed by Zheng et al. (2023) to construct our training pipeline. For all settings, we fine-tuned it for a maximum of 5 epochs, using a single NVIDIA A100-80G GPU. Additionally, we configured the training batch size per device to 1 and established the initial learning rate at 1×10^{-5}, with a linear learning rate scheduler. The best checkpoints were selected based on the performance on the validation data.

You are an English teacher aiming to improve coherence in student writing. You are about to synthesize data for the coherence detection task. Concretely, for each data point, you will be given: a sentence S and a context C, which comprises all preceding sentences up to and immediately before sentence S in an essay written by an English second language learner. Then, you should follow the following steps to create a complete data point:
1) For sentence S and context C, determine if sentence S is coherent with context C. You need to output 1 for [Coherence] if the sentence S is coherent when appended to the context C; otherwise, output 0;
2) Then, if you output 1 in the previous step, output "Done" and finish; otherwise, move on to the following steps;
3) You need to output 1 for [Reason 1] if the sentence S does not connect semantically with the context C; otherwise, output 0;
4) You need to output 1 for [Reason 2] if the new sentence S discusses an entity that has not been introduced in C yet, or the new sentence S discusses an entity that is ambiguous in C; otherwise, output 0;
5) You need to output 1 for [Reason 3] if the relation between sentence S and previous ones in C doesn't make sense due to a missing discourse marker; otherwise, output 0;
6) You need to output 1 for [Reason 4] if the new sentence S contradicts or is inconsistent with previously presented information in C; otherwise, output 0;
7) You need to output 1 for [Reason 5] if the new sentence S introduces information that is completely irrelevant to the context C; otherwise, output 0;
8) You need to output 1 for [Reason 6] if the new sentence S introduces information that is either tangential or slightly irrelevant to the context C; otherwise, output 0;
9) You need to output 1 for [Reason 7] if the comment (rheme/focus) of the sentence does not agree with the topic of the sentence; otherwise, output 0
10) [Rewrite] You should modify sentence S as minimally as possible to improve its coherence based on the following suggestions for each reason you might select above:
- [Reason 1]: add reference words or repeated words or substitutions that can semantically connect sentence S to the context C;
- [Reason 2]: link the newly introduced entity or ambiguous entity in S to the given context C
- [Reason 3]: add or change a discourse marker that ties the sentence S with the given context C
- [Reason 4]: align the newly introduced information with previously introduced information so that the new information in S does not contradict the context C
- [Reason 5]: modify the sentence S so that it is relevant to the context C established by the writer
- [Reason 6]: only output "DELETE" for deleting the sentence S
- [Reason 7]: rewrite sentence S so that the comment of sentence S agrees with the topic of sentence S
Please disregard any incoherences in context C. You should output 1 for [Coherence] only if:
a) sentence S semantically connects to context C, and
b) all entities discussed in the new sentence S have been introduced in C, and
c) sentence S demonstrates reasonable discourse relation with previous ones, and
d) sentence S contains a meaning consistent with previously presented data in C, and
e) sentence S contains a meaning relevant to previously presented data in C.
Here are some examples:
C: I believe that young people nowadays do not give enough time to helping their communities.
S: This, i believe is caused by the environment we live in.
- [Coherence]: 1
- Done
C: Then, I wanna indicate that young people can study many things that are interesting or exciting things for young people.
S: About students, they can learn various fields that students want to study.
- [Coherence]: 0
- [Reason 1]: 1
- [Reason 2]: 0
- [Reason 3]: 1
- [Reason 4]: 0
- [Reason 5]: 0
- [Reason 6]: 0
- [Reason 7]: 0
- [Rewrite]: For example when they study, they can learn various fields that they want to study.
C: There are three main reasons that my ideas support effectively, like action, study and knowledge.
S: First of all, I wanna introduce young people's active points in comparison with older people.
- [Coherence]: 0
- [Reason 1]: 0
- [Reason 2]: 0
- [Reason 3]: 0
- [Reason 4]: 1
- [Reason 5]: 0
- [Reason 6]: 0
- [Reason 7]: 0
- [Rewrite]: First of all, I wanna introduce young people's actions in comparison with older people's.
C: These publicity agents use a lot of techniques to make the products look better, for example they use specialized software like photoshop to increase the size of the product or make it brighter, or maybe an artificial imitation of the product that does not necessarily have the same texture of look.
S: Even though one can observe this situation mostly in food products.
- [Coherence]: 0
- [Reason 1]: 0
- [Reason 2]: 0
- [Reason 3]: 0
- [Reason 4]: 0
- [Reason 5]: 0
- [Reason 6]: 1
- [Reason 7]: 0
- [Rewrite]: DELETE
C: I, however, think in terms of physical and mental factors young people are superior to older people.
S: For example, in the case of sports young people can run and jump, and they can train their muscles that are used in each sport such as transitional sports or silence sports.
- [Coherence]: 0
- [Reason 1]: 0
- [Reason 2]: 0
- [Reason 3]: 0
- [Reason 4]: 0
- [Reason 5]: 0
- [Reason 6]: 0
- [Reason 7]: 1
- [Rewrite]: For example, in the case of sports young people can run and jump, and they can train their muscles for sports more than older people can.
Now, please generate:

Table 6: The prompt for GPT-4 to generate synthetic training data. We provide 4 human-annotated examples in order to constrain its output format.

Label	Cohesion	Consistency	Relevance	Other
Yes	827	511	460	387
No	825	848	848	848

Table 7: Statistics of synthetic training data for the incoherence reasoning task.

F.3 Additional results

The additional results for the incoherence detection and reasoning tasks are shown in Table 8. We can see that training with the DECOR training set D_T usually outperforms training with the out-of-distribution dataset D_C, and training with the combined dataset $D_C + D_T$ can result in a performance uplift, possibly thanks to greater generalizability. A notable point is that for the consistency test set, the labels are relatively imbalanced as mentioned in Section 4, i.e. there are a lot more consistent examples than inconsistent examples, so models that tend to bias towards predicting all examples as consistent would score higher at weighted F1. This could be corrected by using macro F1 or expanding the test set to include more inconsistent examples, which we plan to explore in the future.

G GPT-4 Prompts for Detection, Reasoning, and Rewriting

To leverage the in-context learning capabilities of LLMs, we also prompt GPT-4 in a zero-shot and few (16)-shot setting to establish our baseline results.

G.1 Detection

For coherence detection, Table 10 shows the zero-shot prompt while Table 11 shows the 16-shot prompt.

G.2 Reasoning - Cohesion

For reasoning about the current sentence's cohesion, Table 12 shows the zero-shot prompt while Table 13 shows the 16-shot prompt.

G.3 Reasoning - Consistency

For reasoning about the current sentence's consistency, Table 14 shows the zero-shot prompt while Table 15 shows the 16-shot prompt.

G.4 Reasoning - Relevance

For reasoning about the current sentence's relevance, Table 16 shows the zero-shot prompt while Table 17 shows the 16-shot prompt.

G.5 Reasoning - Others

For reasoning about the current sentence's incoherence that belongs to neither of the above three categories, and instead is a disagreement of the sentence topic with its comment, Table 18 shows the zero-shot prompt while Table 19 shows the 16-shot prompt.

H GPT-4 Judge Prompt

For the incoherence rewriting task, we employ GPT-4 as a judge to conduct the pairwise evaluations and determine which one is better than the other, as explained in Section 5.3. The prompt we use is shown in Table 20.

Models	Training Data	Incoherence Detection (%)	Incoherence Reasoning Selection (%)			
			Cohesion	*Consistency*	*Relevance*	*Others*
Classification-based						
BERT-base	D_C	63.04	48.17	93.76	28.47	-
	D_T	66.43	44.38	75.41	46.37	80.36
	$D_C + D_T$	65.64	47.64	80.40	46.52	-
BERT-large	D_C	64.21	45.93	93.76	28.47	-
	D_T	65.71	44.75	**93.99**	**48.34**	82.65
	$D_C + D_T$	66.26	45.67	**93.99**	42.00	-
DeBERTa-base	D_C	62.21	47.93	93.88	29.45	-
	D_T	**68.54**	46.47	77.17	45.14	74.20
	$D_C + D_T$	67.52	**48.50**	77.97	45.53	-
DeBERTa-large	D_C	53.78	45.93	**93.99**	28.47	-
	D_T	53.78	45.93	92.74	41.36	**89.70**
	$D_C + D_T$	67.05	47.68	84.73	47.02	-
Generation-based						
Llama2-7B	D_C	59.52	43.93	93.65	28.87	-
	D_T	66.08	46.63	83.55	47.20	87.78
	$D_C + D_T$	67.29	43.40	88.26	45.65	-

Table 8: Evaluation of more BERT and DeBERTa models on DECOR using weighted F1 scores for Incoherence Detection and Incoherence Reasoning tasks. Task-specific synthetic data that was used in training is denoted as D_T and out-of-distribution training data from Maimon and Tsarfaty (2023) is denoted as D_C. For each task, we also combine their respective D_C and D_T to train models.

Models	Context	Current Sentence	Reasons	Condition	Rewrite
Llama2-7B	So many investors struggled to make extraordinary shops. And they also have to care about the location of the shop and the menu. Because there are rival shops.	But if someone tries a new thing, he doesn't have to worry about the rival.	Discourse marker	w/ reason	***However***, if someone tries a new thing, he doesn't have to worry about the rival.
				w/o reason	For example, if someone tries a new thing, he doesn't have to worry about the rival.
Llama3-8B -Instruct	I found in several books of scientists, universities magazines, that people who want to succeed need to take risks, risk in research, risk in budgets, contracts, borrow investment.	You can never do big money with the regular risks.	Semantic connection	w/ reason	You can never ***achieve big success*** without taking risks.
				w/o reason	For example, you can never do big business with regular risks.

Table 9: Example rewrites produced by our fine-tuned models, using incoherent context-sentence pairs as input. The reason for the incoherence is also specified. The parts of the sentence that is causing the incoherence are marked as red. Important information from the context is marked with blue. These examples demonstrate that fine-tuning with reasons for incoherence yields better rewrites compared to those produced by the model trained without such reasons.

You are about to perform the task of coherence detection for the sentences written by second-language English learners.
In this task, given a sentence S and a context C, you need to output 1 if S is coherent with C based on the following instructions; otherwise, output 0. You should output 1 only if:
a) sentence S semantically connects to context C, and
b) all entities discussed in the new sentence S have been introduced in C, and
c) the relation between sentence S and previous ones in C makes sense due to proper use of discourse markers, and
d) the new sentence S does not contradict or is not inconsistent with previously presented information in C, and
e) the new sentence S introduces information that is relevant to the context C established by the writer.
Now, please generate:

C: [context]
S: [sentence]

Table 10: Zero-shot prompt for GPT-4 coherence detection.

You are about to perform the task of coherence detection for the sentences written by second-language English learners.
In this task, given a sentence S and a context C, you need to output 1 if S is coherent with C; otherwise, output 0 and provide a concise explanation. Please disregard any incoherences in context C. Specifically, output 0 if:
a) the sentence S does not connect semantically with the context C; or
b) the new sentence S discusses an entity that has not been introduced in C yet, or the new sentence discusses an entity that is ambiguous in C; or
c) the relation between sentence S and previous ones in C doesn't make sense due to an inaccurate discourse marker; or
d) sentence S contradicts or is inconsistent with previously presented information in C; or
e) sentence S introduces information that is completely irrelevant to the context C; or
f) sentence S introduces information that is either tangential or slightly irrelevant to the context C; or
g) the comment of the sentence does not agree with the topic of the sentence itself, or some terms in S are not semantically consistent with each other.
Here are some examples:
C: I believe that young people nowadays do not give enough time to helping their communities.
S: This, i believe is caused by the environment we live in.
Answer: 1
... (14 more examples)
C: I, however, think in terms of physical and mental factors young people are superior to older people.
S: For example, in the case of sports young people can run and jump, and they can train their muscles that are used in each sport such as transitional sports or silence sports.
Answer: 0
Concise explanation: "transitional sports" and "silence sports" are not semantically consistent with each other. They also do not agree with the topic of the sentence.
Now, please generate:

C: [context]
S: [sentence]
Answer:

Table 11: 16-shot prompt for GPT-4 coherence detection.

In this task, given a sentence S that is incoherent with a context C, you need to detect the reason that causes the incoherence. There are seven possible reasons that can cause incoherences:
a) sentence S is incoherent with C because S does not connect semantically with C;
b) sentence S is incoherent with C because S discusses an entity that has not been introduced in C yet, or the new sentence S discusses an entity that is ambiguous in C;
c) sentence S is incoherent with C because the discourse relation between S and previous ones in C doesn't make sense due to an incorrect discourse marker;
d) sentence S is incoherent with C because S contradicts or is inconsistent with previously presented information in C;
e) sentence S is incoherent with C because S introduces information that is completely irrelevant to the context C;
f) sentence S is incoherent with C because S introduces information that is tangential and unnecessary;
g) sentence S is incoherent with C because the comment of the sentence does not agree with the topic of the sentence
In this task, please think step by step and output 1 only if S is incoherent with C due to any of reason a), reason b) or reason c). Otherwise, output 0 if S is incoherent with C due to other reasons. In your answer, start by directly generating either 1 or 0, then followed with reasons. Now, please generate the answer:

C: [context]
S: [sentence]

Table 12: Zero-shot prompt for GPT-4 cohesion reasoning.

In this task, given a sentence S that is incoherent with a context C, you need to detect the reason that causes the incoherence. There are seven possible reasons that can cause incoherences:
a) sentence S is incoherent with C because S does not connect semantically with C;
b) sentence S is incoherent with C because S discusses an entity that has not been introduced in C yet, or the new sentence S discusses an entity that is ambiguous in C;
c) sentence S is incoherent with C because the discourse relation between S and previous ones in C doesn't make sense due to an incorrect discourse marker;
d) sentence S is incoherent with C because S contradicts or is inconsistent with previously presented information in C;
e) sentence S is incoherent with C because S introduces information that is completely irrelevant to the context C;
f) sentence S is incoherent with C because S introduces information that is tangential and unnecessary;
g) sentence S is incoherent with C because the comment of the sentence does not agree with the topic of the sentence
In this task, please think step by step and output 1 only if S is incoherent with C due to any of reason a), reason b) or reason c). Otherwise, output 0 if S is incoherent with C due to other reasons. In your answer, start by directly generating either 1 or 0, then follow with reasons.
Here are some examples:
C: However, the war times have passed and there are fewer who remember or have lived through those conditions and the hardships of life. Now with some people having more money than they actually need there is no strong need to help each other out.
S: Most of us also live in small apartments, where only the father, mother and the child rent.
Answer: 1
Concise explanation: "us" in the sentence does not connect semantically with "people" in the context. Hence, "us" should be changed to "people".
... (14 more examples)
C: As there is a saying that try and try until you succeed and success is the stepping stone this is said because by trying new things only the man can prove himself to be the successful person he must also be confident of what he is doing.
S: If we do the things as we know how to do, it will not cost anything that we cannot gain knowledge of by doing them.
Answer: 0
Concise explanation: This sentence S is tangential and unnecessary.
Now, please generate:

C: [context]
S: [sentence]
Answer:

Table 13: 16-shot prompt for GPT-4 cohesion reasoning.

You are about to perform the task of consistency detection for the sentences written by second-language English learners. In this task, given a sentence S that is incoherent with a context C, you need to output 1 if S contradicts previously presented information in C; otherwise, output 0. Now, please generate the answer:

C: [context]
S: [sentence]
Answer:

Table 14: Zero-shot prompt for GPT-4 consistency reasoning.

You are about to perform the task of consistency detection for the sentences written by second-language English learners. In this task, given a sentence S that is incoherent with a context C, you need to output 1 if S contradicts previously presented information in C; otherwise, output 0.
Here are some examples:
C: Then, I wanna indicate that young people can study many things that are interesting or exciting things for young people.
S: About students, they can learn various fields that students want to study.
Answer: 0
Concise explanation: Sentence S does not contradict previously presented information in C. S is incoherent with C because "students" does not connect semantically with "young people" in the context.
... (14 more examples)
C: As there is a saying that try and try until you succeed and success is the stepping stone this is said because by trying new things only the man can prove himself to be the successful person he must also be confident of what he is doing.
S: If we do the things as we know how to do, it will not cost anything that we cannot gain knowledge of by doing them.
Answer: 0
Concise explanation: Sentence S does not contradict previously presented information in C. This sentence S is tangential and unnecessary.
Now, please generate:

C: [context]
S: [sentence]
Answer:

Table 15: 16-shot prompt for GPT-4 consistency reasoning.

In this task, given a sentence S that is incoherent with a context C, you need to output 1 if S is incoherent with C because of a lack of relevance based on the following instructions; otherwise, output 0. You should output 1 only if:
a) sentence S introduces information that is completely irrelevant to context C, or
b) sentence S introduces information that is either tangential or slightly irrelevant to context C.

C: [context]
S: [sentence]
Answer:

Table 16: Zero-shot prompt for GPT-4 relevance reasoning.

In this task, given a sentence S that is incoherent with a context C, you need to output 1 if S is incoherent with C because of a lack of relevance based on the following instructions; otherwise, output 0. You should output 1 only if:
a) sentence S introduces information that is completely irrelevant to context C, or
b) sentence S introduces information that is either tangential or slightly irrelevant to context C.
Here are some examples:
C: Then, I wanna indicate that young people can study many things that are interesting or exciting things for young people.
S: About students, they can learn various fields that students want to study.
Answer: 0
Concise explanation: Sentence S introduces information that is relevant to context C. S is incoherent with C because "students" does not connect semantically with "young people" in the context.
... (14 more examples)
C: As there is a saying that try and try until you succeed and success is the stepping stone this is said because by trying new things only the man can prove himself to be the successful person he must also be confident of what he is doing.
S: If we do the things as we know how to do, it will not cost anything that we cannot gain knowledge of by doing them.
Answer: 1
Concise explanation: Sentence S introduces information that is irrelevant and tangential to context C. This sentence S is tangential and unnecessary. Hence, Sentence S is incoherent with context C because of the relevance issue.
Now, please generate:

C: [context]
S: [sentence]
Answer:

Table 17: 16-shot prompt for GPT-4 relevance reasoning.

In this task, given a sentence S that is incoherent with a context C, you need to output 1 if S is incoherent with C because of the disagreement between the topic and the comment of sentence S; otherwise, output 0. Specifically, you should output 1 only if the comment of sentence S does not agree with the topic of the sentence itself. Now, please generate the answer:

C: [context]
S: [sentence]
Answer:

Table 18: Zero-shot prompt for GPT-4 reasoning for other categories, e.g. topic-comment disagreement.

In this task, given a sentence S that is incoherent with a context C, you need to output 1 if S is incoherent with C because of the disagreement between the topic and the comment of sentence S; otherwise, output 0. Specifically, you should output 1 only if the comment of sentence S does not agree with the topic of the sentence itself.
Here are some examples:
C: I, however, think in terms of physical and mental factors young people are superior to older people.
S: For example, in the case of sports young people can run and jump, and they can train their muscles that are used in each sport such as transitional sports or silence sports.
Answer: 1
Concise explanation: The comment of sentence S does not agree with the topic of the sentence. "transitional sports" and "silence sports" are not consistent with each other in sentence S itself, and they also do not agree with the topic of the sentence. Hence, Sentence S is incoherent with context C because of the disagreement between the topic and the comment of sentence S.
... (14 more examples)
C: As there is a saying that try and try until you succeed and success is the stepping stone this is said because by trying new things only the man can prove himself to be the successful person he must also be confident of what he is doing.
S: If we do the things as we know how to do, it will not cost anything that we cannot gain knowledge of by doing them.
Answer: 0
Concise explanation: The comment of sentence S agrees with the topic of the sentence. This sentence S is tangential and unnecessary. Hence, Sentence S is incoherent with context C because of the relevance issue.
Now, please generate:

C: [context]
S: [sentence]
Answer:

Table 19: 16-shot prompt for GPT-4 reasoning for other categories, e.g. topic-comment disagreement.

You are an English teacher aiming to improve coherence in student writing.
You need to evaluate and select the best system based on the coherence of their outputs to a given instruction.
This process will be used to create a leaderboard reflecting the most accurate and human-preferred answers.
I require a leaderboard for various systems. I'll provide you with prompts given to these models and their corresponding outputs.
Your task is to assess these responses, and select the model that produces the most coherent output from a English teacher's perspective.

Instruction

You are about to perform the task of sentence rewriting for the sentences written by second-language English learners.
In this task, given a context C and a sentence S, where S is incoherent with C, you need to rewrite sentence S to make it coherent with C according to the following instructions: The rewrite should be as minimal as possible. [reason_texts].
C: [context]
S: [sentence]
Answer:

Model Outputs

Here are the unordered outputs from the models. Each output is associated with a specific model, identified by a unique model identifier.

model_identifier: "m"
output: [output_1]

model_identifier: "M"
output: [output_2]

Task

Evaluate the models based on the coherence of their outputs to the given context C, and select the model that generated the most coherent output.
The output from the model should rewrite the incoherent sentence S as minimally as possible.
Retaining awkward phrasing or minor grammar errors from sentence S is acceptable as long as the output is coherent with context C.
Answer by first providing a concise explanation and then end your answer by providing the model identifier of the most coherent output.
We will use the last character of your output 'output[-1]' as the name of the best model, so make sure you finish with the token of the model identifiers and nothing else: 'm' or 'M' (no quotes, no dots, no backticks, no new lines, ...). For example:

Concise explanation
...some text...

Which is best, m or M?
1

Now is your turn.

Your answer: "Concise explanation" followed by "Which is best, m or M?"

Table 20: The prompts for GPT-4 as a judge to conduct the pairwise comparisons between model outputs.