Detecting AI-Generated Sentences in Human-AI Collaborative Hybrid Texts: Challenges, Strategies, and Insights

Zijie Zeng¹, Shiqi Liu¹,², Lele Sha¹, Zhuang Li¹, Kaixun Yang¹
Sannyuya Liu², Dragan Gašević¹ and Guanliang Chen¹,*

¹Centre for Learning Analytics, Monash University, Australia
²National Engineering Research Center of Educational Big Data, Central China Normal University, China
{zijie.zeng, lele.sha, zhuang.li, kaixun.yang1, dragan.gasevic, guanliang.chen}@monash.edu
liushiqi@mails.ccnu.edu.cn, liusy027@ccnu.edu.cn

Abstract

This study explores the challenge of sentence-level AI-generated text detection within human-AI collaborative hybrid texts (abbreviated as hybrid texts). Existing studies of AI-generated text detection for hybrid texts often rely on synthetic datasets. These typically involve hybrid texts with a limited number of boundaries, e.g., single-boundary hybrid texts that begin with human-written content and end with machine-generated continuations. We contend that studies of detecting AI-generated content within hybrid texts should cover different types of hybrid texts generated in realistic settings to better inform real-world applications. Therefore, our study utilizes the CoAuthor dataset, which includes diverse, realistic hybrid texts generated through the collaboration between human writers and an intelligent writing system in multi-turn interactions. We adopt a two-step, segmentation-based pipeline: (i) detect segments within a given hybrid text where each segment contains sentences of consistent authorship, and (ii) classify the authorship of each identified segment. Our empirical findings highlight (1) detecting AI-generated sentences in hybrid texts is overall a challenging task because (1.1) human writers’ selecting and even editing AI-generated sentences based on personal preferences adds difficulty in identifying the authorship of segments; (1.2) the frequent change of authorship between neighboring sentences within the hybrid text creates difficulties for segment detectors in identifying authorship-consistent segments; (1.3) the short length of text segments within hybrid texts provides limited stylistic cues for reliable authorship determination; (2) before embarking on the detection process, it is beneficial to assess the average length of segments within the hybrid text. This assessment aids in deciding whether (2.1) to employ a text segmentation-based strategy for hybrid texts with longer segments, or (2.2) to adopt a direct sentence-by-sentence classification strategy for those with shorter segments.

1 Introduction

AI-generated text detection refers to the process of identifying the text generated by AI systems. The significance of this task lies in its dual role as both a deterrent and a defense against the potential misuse of modern, powerful generative AI technologies, particularly in educational contexts. According to [Ma et al., 2023; Zellers et al., 2019], contemporary advanced language models are capable of producing deceptive content designed to manipulate public perception. This includes the creation of fictitious news [Zellers et al., 2019], counterfeit application reviews [Martens and Maalej, 2019], and fabricated social media posts [Fagni et al., 2021]. Additionally, education practitioners have expressed concerns about the growing popularity of the use of ChatGPT (a generative AI system developed by OpenAI) among students. They emphasize that over-reliance on ChatGPT could hinder the development of students’ writing and critical thinking skills [Ma et al., 2023; Mitchell et al., 2023] and, in the worst cases, lead to breaches of academic integrity [Currie, 2023; Sun and Hoelscher, 2023].

Existing studies [Koike et al., 2024; Hu et al., 2024; Ma et al., 2023; Mitchell et al., 2023; Clark et al., 2021; Jawahar et al., 2020; Martens and Maalej, 2019] have primarily focused on detecting AI-generated text at the granularity of paragraphs or documents, operating under the assumption that the text under scrutiny is either entirely AI-generated or not at all. However, the framing of the detection task as a binary classification problem at the document level has recently been challenged by [Dugan et al., 2023], who have observed the trends in human-AI collaborative writing [Buschek et al., 2021; Lee et al., 2022]. They noted that a piece of text (e.g., a document) could contain both human-written and AI-generated content. This observation underscores the need for detecting AI-generated text at a finer granularity level (e.g., at the sentence level instead of the document level). Although pilot studies [Wang et al., 2023; Zeng et al., 2024] have been conducted to detect AI-generated content within hybrid texts, we are concerned that the strategies these studies used to synthesize hybrid texts are overly simplistic and lack variation. For example, a simple practice of synthesizing a hybrid text (as adopted by [Wang et al., 2023; Zeng et al., 2024]) involves using a piece of human-written text as a starting point, and then instructing a generative
AI tool (e.g., ChatGPT) to write a continuation based on the given human-written beginning text, resulting in a synthesized hybrid text with only one boundary. Behind this concern lies our doubt regarding whether a detection strategy or approach that works well on relatively simple hybrid texts (e.g., those with fewer boundaries such as the single-boundary ones) can be equally effective on complex hybrid texts, e.g., hybrid texts with up to dozens of boundaries that are generated through multi-turn interactions between human writers and intellectual writing assistants. To investigate this, we posit that a comprehensive study of AI-generated text detection from hybrid texts should be grounded on a dataset that encompasses hybrid texts generated with varying human-AI interaction styles and strategies. Therefore, in this study, we conducted further investigation to detect AI-generated text within hybrid texts. It is noteworthy that these hybrid texts originate from the CoAuthor dataset [Lee et al., 2022] and are collaboratively created by human writers and intelligent writing assistants through a series of interactions in real-world environments. We formally define our Research Question as:

- To what extent can AI-generated sentences be automatically detected in hybrid texts co-created by human writers and intelligent writing assistants?

Recognizing the fact that a hybrid text can be divided into segments, with each segment consisting of several sentences that share the same authorship, we use a two-step segmentation-based pipeline [Tepper et al., 2012; Xia and Wang, 2023] (i.e., segment detection followed by segment classification) for sentence-level AI-generated text detection from hybrid texts. However, as noted in [Zeng et al., 2024], as the number of boundaries increases, segment detection approaches tend to become less accurate in identifying boundaries between human-written and machine-generated segments. Thus, we also introduce a naive segmentation strategy, treating each sentence as an individual segment, to facilitate comparisons between text segmentation-based strategies and the straightforward sentence-by-sentence classification strategy. Through extensive empirical experiments, we have summarized the following main findings: (1) detecting AI-generated sentences within hybrid texts is overall a challenging task because (1.1) the complex interplay between human and AI writing systems, where humans could select and even edit AI-generated sentences based on personal preferences, adds difficulty in identifying the authorship of segments; (1.2) the frequent change of authorship between neighboring sentences within the hybrid text creates difficulties for segment detectors in accurately identifying authorship-consistent segments, thereby hampering the downstream authorship identification task; (1.3) the short length of text segments offering limited stylistic information for accurate authorship identification; (2) with a sufficiently high-performing segment detection model and segment classification model, the two-step segmentation-based pipeline solution may perform better than approaches that jointly learn these two tasks; (3) Before starting the detection task, assessing (through prior knowledge or other machine learning techniques) the average segment length of hybrid texts is useful, as it can influence the strategy of either: (3.1) using the text segmentation-based strategy for hybrid text with longer segments; or (3.2) opting for the direct sentence-by-sentence classification strategy for hybrid text with shorter segments.

2 Related Work

Al-Generated Text Detection. As modern LLMs’ generative capabilities strengthen, distinguishing between AI-generated texts and human-written texts becomes increasingly challenging [Ma et al., 2023]. This may lead to the misuse of generative AI in scenarios where it should not be used, such as students employing it for writing assignments, which can result in academic misconduct [Mitchell et al., 2023]. These potential risks emphasize the necessity for effective detection of AI-generated text. Existing studies [Koike et al., 2024; Hu et al., 2024; He et al., 2023; Mitchell et al., 2023; Pagnoni et al., 2022; Rosati, 2022] have been criticized for their narrow focus on document-level detection and the assumption that a document is either entirely AI-generated or not. This point is noted in [Dugan et al., 2023], which suggests that AI-generated text detection should be conducted with finer granularity to adapt to the increasingly popular trend of human-AI collaborative writing. Specifically, [Zeng et al., 2024] introduced a boundary detector for detecting boundaries between human-written sentences and AI-generated sentences. Meanwhile, [Wang et al., 2023] adopted a more straightforward approach to detection, namely, identifying the exact authorship of each sentence within the hybrid text. They proposed a method called SeqXGPT to extract word-wise log probability lists and use a combination of CNN and Transformer for sentence-level authorship detection within hybrid texts.

Text Segmentation. Text segmentation aims to discover the latent structure of a document by dividing the text into segments with different labels [Ghinassi et al., 2023; Bai et al., 2023; Lo et al., 2021; Xia and Wang, 2023] (e.g., segmentation based on different topics). The broad spectrum of text segmentation studies can be further divided into two categories: (1) those focusing solely on boundary detection for segmentation [Zeng et al., 2024; Lukasik et al., 2020; Ghinassi et al., 2023; Yu et al., 2023; Xing et al., 2020; Li et al., 2022; Somasundaran and others, 2020; Koshorek et al., 2018]; and (2) those aiming at both segment detection and subsequent segment classification [Bai et al., 2023; Lo et al., 2021; Xia and Wang, 2023; Gong et al., 2022; Tepper et al., 2012]. Among the many possible approaches for segmentation, we have selected three approaches with open-source code for our segment detection module: (i) TriBERT [Zeng et al., 2024], because it was specifically developed for distinguishing AI-generated segments and human-written  

\[ \text{average segment length} = \frac{h_1 + h_2 + h_3 + m_4}{4} \]

The concept of average segment length can be illustrated with the following two hybrid texts \( T = < h_1, h_2, h_3, m_4 > \) and \( R = < h_5, m_6, h_7, m_8 > \), where \( h_i \) and \( m_i \) are human-written sentence and machine-generated sentence, respectively. The average segment length \( l_T \) for \( T \) can be calculated in a weighted manner as \( l_T = (3/4 \times 3) + (1/4 \times 1) \times 1 = 1 < 2.5 \), i.e., a shorter average length reflects the more frequent change of authorship between neighboring sentences. Note that we assume each sentence has a length of 1.
segments within hybrid texts; (ii) Transformer \cite{Lo et al., 2021}, due to its widespread adoption as a baseline in existing text segmentation studies \cite{Xia et al., 2022; Xia and Wang, 2023; Bai et al., 2023; Yu et al., 2023}; and (iii) SegFormer \cite{Bai et al., 2023}, chosen as a representative of the state-of-the-art (also one of the latest) text segmentation approaches.

**Human-AI Collaborative Writing and Datasets.** Modern large language models have made human-AI collaborative writing increasingly convenient. Existing studies show that the suggestions provided by intelligent writing assistants have evolved from phrase-level \cite{Buschek et al., 2021} to sentence-level \cite{Lee et al., 2022}. Furthermore, these assistants seem poised for further improvement, especially with the release of GPT-4. Pilot studies \cite{Dugan et al., 2023; Zeng et al., 2024; Wang et al., 2023} have been conducted to address AI-generated text detection within hybrid text data at a granularity finer than the document level. However, these studies were mainly conducted on synthetic and simplistic datasets due to their low cost and ease of acquisition. For instance, \cite{Wang et al., 2023} instructed the GPT-3.5-turbo API to generate continuations on given human-written texts. \cite{Zeng et al., 2024} goes further by designing a set of six prompting templates that enable ChatGPT to generate multi-boundary hybrid texts (no more than three boundaries) from incomplete student-written essays. For each of these original essays, specific segments were removed and substituted with new sentences generated by ChatGPT. However, these synthetic texts are still overly simplistic compared to the hybrid texts generated through multi-turn human-AI interactions in realistic scenarios, which could involve up to dozens of boundaries. We contend that to guide and enlighten real-world practices, AI-generated text detection studies should not rely solely on synthetic datasets. Therefore, our study aims at the CoAuthor dataset \cite{Lee et al., 2022}, which includes diverse, realistic hybrid texts generated through human-AI interactions and can serve as an ideal benchmark for exploring AI-generated text detection in hybrid texts.

### 3 Methods

#### 3.1 Dataset and Task

**Dataset.** We selected the CoAuthor Dataset \cite{Lee et al., 2022}, which consists of texts generated through multi-turn interactions between human writers and a GPT-3-based intelligent writing assistant. The dataset includes 1,445 essays from 63 writers who interacted with the writing assistant across multiple writing sessions. Writers were given a prompt text as a starter and could either write independently or request sentence suggestions from the assistant. These suggestions could be accepted (with or without edits) or rejected, and this process continued until an essay was completed\(^2\).

**Task.** For a hybrid text \(H\) comprising \(n\) sentences, represented as a sequence of sentences \(\{s_1, s_2, ..., s_n\}\), an algorithm is required to predict a list of labels \(\{l_1, l_2, ..., l_n\}\) for \(H\). Each label \(l_i\) corresponds to the predicted authorship for sentence \(s_i\), e.g., in CoAuthor dataset, \(l_i\) could be AI-generated, Human-written, or Human-AI Collaborative.

### 3.2 Approaches

We adopt a two-step segmentation-based pipeline to address the sentence-level AI-generated text detection from hybrid texts, involving segment detection followed by segment classification \cite{Tepper et al., 2012; Xia and Wang, 2023}. This pipeline approach is based on our observation that a hybrid text can be divided into segments, with each segment consisting of several sentences that share the same authorship. We begin by introducing the **segment detection module:**

- **Perfect Segment Detector:** This approach serves as an ideal segment detector, predicting all boundaries with 100% accuracy.

- **Naïve Segment Detector:** This approach treats each sentence in the hybrid text as a separate segment, which is essentially equivalent to performing direct sentence-by-sentence classification over the hybrid text.

- **TriBERT:** TriBERT \cite{Zeng et al., 2024} is designed to identify transitions between segments of differing authorship (e.g., human vs. AI-generated text). It measures distances between adjacent segments, pinpointing transitions where this distance is greatest. However, it only ranks boundary candidates by their probabilities of being true boundaries. Therefore, we manually selected the top \(p\) boundaries from the sorted candidate list with \(p \in \{20\%, 40\%, 60\%, 80\%\}\).

- **SegFormer:** SegFormer \cite{Bai et al., 2023} is a topic segmentation approach that utilizes unidirectional attention blocks for enhanced sentence modeling. It incorporates a specialized context aggregator and leverages topic classification loss to reduce noise and enhance information processing. Note that SegFormer is capable of not only identifying boundaries between segments but also providing a predicted label for each segment.

- **Transformer**\(^2\) \cite{Lo et al., 2021}: This approach merges bottom-level pre-trained transformers for sentence embedding with an upper-level transformer to identify segments and predict their labels (i.e., authorship).

Then we present the approaches for the **segment classification module** as follows:

- **BERT:** This method utilizes the BERT-base \cite{Kenton and Toutanova, 2019} model to encode the textual input into an embedding. This embedding is then fed into a dense classification layer for label prediction.

- **SeqXGPT:** SeqXGPT \cite{Wang et al., 2023} utilizes the log probability lists generated by transparent LLMs as its primary features. Considering that these features mirror the wave patterns observed in speech processing and surpass the analytical capabilities of LLMs, it further integrates convolutional and self-attention networks into its core architecture to process these features.

\(^2\)TriBERT with \(p = 100\%\) is equivalent to the Naïve segment detector, i.e., the sentence-by-sentence classification strategy.
DeBERTa-v3: DeBERTa-v3 [He et al., 2023] is a recently developed language model that has set new benchmarks in several natural language processing (NLP) tasks. This advanced version builds on its predecessor, DeBERTa [He et al., 2020], by implementing an enhanced pre-training task with improved sample efficiency called replaced token detection (RTD), instead of the traditional masked language modeling task.

DistilBERT: This method resembles the above method, BERT, but it adopts an alternative encoder, namely DistilBERT [Sanh et al., 2019], which is a distilled, efficient version of BERT.

RoBERTa: This method resembles the above method, BERT, but it adopts RoBERTa [Liu et al., 2019] as the encoder, which can be viewed as an optimized version of BERT, with changes primarily in the training regime and data size.

GPT-3.5 (Fine-tuned): We first format our training data as follows: Input: <TEXT SEGMENT>, Output: <LABEL>. Then, we fine-tune the OpenAI GPT-3.5-Turbo-1106 API using the above-formatted training data and obtain our AI-generated text classifier.

GPT-2: This model adopts the GPT-2 [Radford et al., 2019] model with a classification head on top of the hidden states of the last token for classification.

BERT (Token), DistilBERT (Token), and RoBERTa (Token): This series of methods classifies each token in a text segment, then determines the segment’s label based on token-level predictions. If all tokens are predicted as AI-generated or human-written, the segment is labeled accordingly. If both types are present, the segment is considered as human-AI collaborative. We use BERT, DistilBERT, and RoBERTa as encoders for token embeddings, followed by a classification layer for token-level authorship prediction. These methods are termed BERT (Token), DistilBERT (Token), and RoBERTa (Token).

Transformer and SegFormer: As mentioned above, these two methods can identify boundaries for text segmentation as well as predict labels for the identified segments. Therefore, we also consider Transformer and SegFormer as segment classification models.

3.3 Experiment Setting

Data Splitting. We randomly divided the 1,445 hybrid texts in the CoAuthor dataset into training, validation, and testing sets in a ratio of 70%-15%-15%. Note that the above split datasets were used for the training, validation, and testing processes of all models, including segment detection models and segment classification models.

Evaluation Metric. We observed that accurately detecting segments within a hybrid text does not ensure the precise identification of the authorship for these detected segments. For example, in a hybrid text comprising four sentences < s₁, s₂, s₃, s₄ >, if a segment detector correctly identifies the boundary between sentences s₂ and s₃, it remains ambiguous whether the segment < s₁, s₂ > (or < s₃, s₄ >) is AI-generated or human-written. The segment detector merely indicates a transition in authorship, without specifying which segment is AI-generated. Consequently, we opted for the Kappa score as our evaluation metric, which measures the inter-rater agreement between lists of predicted and actual authorship labels, rather than the Pₚ metric, commonly adopted for evaluating text segmentation performance [Beeferman et al., 1999; Bai et al., 2023; Lo et al., 2021].

Model Training. As training details vary across different models, we itemize them as follows:

TriBERT: We adhere to its default setting [Zeng et al., 2024] to evaluate the model’s performance on the validation set after each training epoch. If the validation performance at epoch t + 1 is not as good as at epoch t, we stop the training and use the best model (saved at epoch t) as the segment detector for the test stage.

Naive Segment Detector and Perfect Segment Detector: These two models require no training process. The first model treats each sentence as a separate segment, while the second model directly splits the hybrid texts into segments using the ground-truth boundaries.

SegFormer and Transformer: These two approaches consider the whole piece of hybrid text as the minimum input and define an epoch as a single iteration of training across all hybrid texts in the training set.

Segment Classification Models: For all segment classification models, a training sample refers to a single sentence. During the test phase, the authorship of a segment is determined based on the majority authorship of its sentences, as predicted from a list containing the authorship for each sentence within the segment.

Important Parameters and Other Details. For all models, the learning rate was selected from {2e−5, 5e−5, 1e−4}. We used a small batch size of 4 for SegFormer due to its higher memory requirements. For other approaches, we consistently adopted a batch size of 32. All tests were conducted using an NVIDIA Tesla A40 GPU equipped with 48 GB of RAM. For details of other model-specific parameters, please refer to our codes available on GitHub.

4 Results

We present our empirical results in Table 1. This includes Sub-table G0, which reports the Kappa score (a measure of classification accuracy) performance for the entire set of hybrid text samples in the test set. Additionally, Sub-tables G1, G2, and G3 display the performance (Kappa score) for different groups of hybrid texts in the test set. The groups have been categorized based on the average segment length (denoted as l) of the hybrid texts, i.e., l ≤ 3 in G1; l ∈ (3, 6] in
Table 1: This table presents empirical findings on Kappa score performance for hybrid text samples, with sub-tables G0 to G3 showing performances for different text groups categorized by average segment length: G1 with \( l < 3 \), G2 with \( 3 \leq l < 6 \), and G3 with \( l \geq 6 \). Enhancements for readability include numbered rows and columns, and bolded values indicating top performance (excluding Perfect) in each row. Note that SegFormer and Transformer\( ^2 \) can serve as either segment detectors (results in Column C6 and C7) or directly as segment classifiers (results in Row R12 and R13). Each reported entry is a mean over three independent runs.

\[ G2; \text{and } l \geq 6 \text{ in } G3. \text{ This categorization enables our investigation into whether the choice of the optimal detection strategy changes according to variations in the average segment length of the hybrid texts. It is noted that a shorter average segment length often indicates more frequent changes in authorship between neighboring sentences, leading to a higher number of boundaries (see Table 2).} \]

Which is the Best Segment Classifier? By referring to the C5 column (i.e., results of the Naïve segment detector) of any one of the sub-tables, we can observe that DeBERTa-v3,
BERT, SeqXGPT, RoBERTa, and DistilBERT outperform the other segment classifiers, with DeBERTa-v3 being the most effective. Additionally, we notice that the token-classification based methods, namely BERT (Token), DistilBERT (Token), and RoBERTa (Token), consistently underperform compared to their sentence-level classification counterparts, which are BERT, DistilBERT, and RoBERTa, demonstrating the effectiveness of the strategy where a BERT-based encoder is directly combined with a classification layer for sentence-level text classification.

**Which is Better? Joint Learning Approaches vs. Two-Step Pipeline Approaches.** As previously mentioned, SegFormer and Transformer are approaches that jointly learn the two tasks of segment detection and segment classification and can serve as either segment detectors or segment classifiers. From our empirical results, we observed an interesting phenomenon: when these two approaches act solely as segment detectors and are combined with high-performing downstream classifiers (e.g., DeBERTa-v3), they may achieve more effective outcomes than when serving directly as segment classifiers. For example, in sub-tables G2 and G3, the SegFormer (or Transformer) and DeBERTa-v3 pipeline (see < R1, C6 > and < R1, C7 >) outperformed instances where SegFormer (or Transformer) alone served as the segment classifier (R12 and R13). These observations suggest that a pipeline with high-performing segment detectors and high-performing segment classifiers does not necessarily perform worse than approaches that jointly learn these two tasks.

**How to Choose the Right Strategy? Sentence-by-Sentence Classification vs. Text Segmentation-Based Classification.** We notice that the best segment detector varies across groups with different average segment length. Specifically, for group G1 with an average segment length of 3 or less, the Naive segment detector achieved the highest performance (averaged across all classifiers), as indicated by a Kappa score of 0.3070 in < R11, C5 >: For group G2, although the Naive segment detector remains the best in terms of mean performance (see < R11, C5 > of the sub-table for G2), we observe a notable difference from the results in group G1, i.e., only 5 out of the 10 top performances (highlighted in bold) originate from column C5; Finally, turning to the sub-table for group G3, we can see from Row R11 that the Naive segment detector performed the worst, while the SegFormer detector achieved the best mean performance.

We first explain why the three realistic segment detectors (i.e., SegFormer, Transformer, and TriBERT) significantly underperform compared to the Naive segment detector (i.e., sentence-by-sentence classification strategy) in group G1 (see sub-table G1, Row R11). We posit that the reason lies in the fact that hybrid texts in group G1 have the shortest average segment length, which typically means a higher number of boundaries (see Table 2). A greater number of boundaries implies increased difficulty in detecting all true boundaries. **Failing to detect all true boundaries** inevitably leads to the creation of authorship-inconsistent segments, i.e., sentences with differing authorships are incorrectly grouped into a single segment (e.g., segments e and f in Figure 1) by an imperfect segment detector. The sentences within these authorship-inconsistent segments, in theory, cannot be predicted with 100% accuracy by downstream segment classifiers, thereby compromising the effectiveness of sentence-level AI-generated text detection. We provide an example in Figure 1 to illustrate that, given the number of sentences being roughly equal in all hybrid texts (as is the case for the CoAuthor dataset), an imperfect segment detector is more likely to produce authorship-inconsistent segments in hybrid texts with a higher number of boundaries than in those with fewer boundaries: (i) Sub-figure (1) illustrates a hybrid text (with only one true boundary), from which an imperfect segment detector identifies one true boundary and two false boundaries, consequently dividing this text into four segments: a, b, c, and d; (ii) Sub-figure (2) displays a hybrid text with five boundaries and an imperfect segment detector splits this hybrid text into segments e, f, g, and h based on the three identified boundaries (one is true and the other two are false). In this example, an imperfect segment detector identified three boundaries in two hybrid texts, each with one true and two false boundaries, resulting in a precision of 1/3. The recall rates for true boundaries were 1.0 in the left-side text and 0.2 in the right-side text, indicating that the detector missed 80% of true boundaries in the right-side text. As mentioned earlier, failing to detect the true boundaries will inevitably result in authorship-inconsistent segments. Take segments e and f in Sub-figure (2) as examples. Due to the mixed-authorship nature of segments e and f, if e is predicted to be AI-generated (or human-written), then the human-written (or AI-generated) sentences within...

---

### Table 2: This table details the ratio of segments with inconsistent authorship across different groups of hybrid texts from the test set.

<table>
<thead>
<tr>
<th>Average Segment Length l within Group</th>
<th>Average Number of Boundaries within Group</th>
<th>Ratio of the Group’s Samples to the Total</th>
<th>Ratio of Segments with Inconsistent Authorship</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1: l &lt;= 3</td>
<td>13.7</td>
<td>46.7%</td>
<td>36.30%</td>
</tr>
<tr>
<td>G2: 3 &lt; l &lt;= 6</td>
<td>9.9</td>
<td>27.9%</td>
<td>29.70%</td>
</tr>
<tr>
<td>G3: l &gt; 6</td>
<td>4.2</td>
<td>25.3%</td>
<td>22.90%</td>
</tr>
</tbody>
</table>

---

8Realistic segment detectors other than Perfect and Naive.
this segment will (inevitably and incorrectly) be labeled as AI-generated (or human-written). This is because the downstream classifier will only predict a single authorship for each identified segment. We further define $r$ as the ratio of segments with inconsistent authorship in a hybrid text $ht$ from the test set, using a segment detector (denoted as $dt$) like TriBERT, SegFormer, or Transformer$^2$. We calculate $r$ for each $<ht, dt>$ pair and then average these values within groups $G1$, $G2$, and $G3$ (see Table 2). The results show $G1$ with the highest inconsistency in authorship, followed by $G2$, and $G3$ with the least, supporting our assumptions as seen in Figure 1. Finally, it should be noted that (i) the Perfect detector identifies all true boundaries without error; (ii) the Naive segment detector, which considers each sentence a separate segment, does not miss true boundaries but detects false ones. Interestingly, as shown in Row R11 of Sub-Table G1, the Naive detector outperforms other detectors, suggesting that the negative impact of false boundaries is less significant than missed boundaries. For example, in Sub-Figure (1), segments $a$ and $b$, originally part of a larger segment $ab$, are split but remain authorship-consistent. Thus, a segment classifier could still theoretically achieve 100% accuracy on $a$ and $b$. However, this unnecessary splitting leads to the challenge of short-text classification [Wang et al., 2018; Guo et al., 2023; Wang et al., 2021], e.g., predicting separately for $a$ and $b$ tends to be less accurate than predicting for the combined segment $ab$.

The analysis above could also shed light on why the three realistic segment detectors (i.e., SegFormer, Transformer$^2$, and TriBERT) outperform the sentence-by-sentence classification strategy (i.e., Naive segment detector) in Group $G3$ (Please refer to Row R11, sub-table $G3$, for the exact performance results.). Specifically, we have the following explanation: the average segment length in Group $G3$ is relatively longer and consequently with fewer boundaries, so that (1) the chances of the true boundaries being missed are rarer, resulting in the lower ratio of authorship-inconsistent segments in $G3$ than in $G1$, as can be seen from the far right column of Table 2; and (2) segments with longer length can help alleviate the short-text classification challenge faced by the downstream segment classifier.

Is Sentence-Level AI-generated Text Detection from Hybrid Texts a Challenging Task? The results in column $C9$, within sub-table $G0$ (where each cell represents the average performance of a specific segment classifier across all segment detection models), indicate that the performance of all included segment classifiers is generally unsatisfactory. The highest Kappa score recorded is 0.3447, achieved by DeBERTa-v3. Column $C8$ reveals that pipelines with a Perfect segment detector significantly outperform others, yet the top Kappa score is only 0.5166, achieved by combining a Perfect segment detector with DeBERTa-v3. We posit that the challenging nature of this task can be attributed to the following aspects: (1) firstly, the difficulty could be explained by looking into the human-AI collaboration process [Lee et al., 2022] which resulted in the CoAuthor hybrid texts: (1.1) each machine-generated sentence in these hybrid texts was selected (or preferred) by a human writer from a set of suggested sentences, which could cause the chosen generated sentence to deviate from the original distribution of machine-generated sentences, aligning more closely with the distribution of human-written sentences. (1.2) human writers could make edits to the selected sentences, resulting in a third class of sentence (the human-AI collaborative sentence), which is in addition to the purely human-written sentences and the purely AI-generated sentences; (2) secondly, the frequent change of authorship between neighboring sentences within the hybrid text creates difficulties for realistic (imperfect) segment detectors in accurately identifying authorship-consistent segments, compromising the effectiveness of the subsequent authorship identification task, as has been discussed in the previous subsection; (3) lastly, even when the segments are perfectly detected (i.e., Column $C8$), identifying the authorship of these segments remains difficult due to the short length of the segments. As can be seen from Table 2, nearly 50% of the hybrid texts in the test set have an average segment length of 3 or less.

5 Conclusion This study investigates the recently popular task of AI-generated text detection within human-AI collaborative hybrid texts. Differing from existing studies [Wang et al., 2023; Dugan et al., 2023; Zeng et al., 2024] that primarily rely on synthetic and simplistic datasets, we have grounded our research on the CoAuthor dataset [Lee et al., 2022], which encompasses a wide range of diverse and realistic hybrid texts, produced through collaborative efforts between human writers and an intelligent writing system during multi-turn interactions. Although our experimental results suggest that this task is generally challenging, we have distilled some practical tips from our findings to assist potential practitioners and researchers facing this task: (1) with a sufficiently high-performing segment detector and segment classifier, the two-step pipeline that combines text segmentation and segment authorship identification could be more reliable than approaches that jointly learn these two tasks; (2) before embarking on the detection process, it is beneficial to assess (through prior knowledge or other machine learning techniques) the average length of segments (or the number of boundaries) within the hybrid texts. This assessment aids in deciding whether to employ a text segmentation-based strategy for hybrid texts with longer segments (usually having fewer boundaries) or to adopt a direct sentence-by-sentence classification strategy for texts with shorter segments, which often have more boundaries. An application case of the above tips would be to employ a text segmentation-based strategy on hybrid texts known to contain only very few boundaries. This scenario can occur when students aim to spend only the least effort on a writing assignment, e.g., preferring to have a generative LLM finish the assignment from a given start text (resulting in single-boundary hybrid text) rather than engaging in more time-consuming and effort-intensive multi-turn interactions. Moreover, considering that the average segment length of hybrid texts influences the choice of the optimal segment detection strategy, exploring whether the average segment length of a given hybrid text can be accurately predicted represents a promising direction for future research efforts.
References


[Hu et al., 2024] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In Proceedings of the 38th AAAI Conference on Artificial Intelligence, Vancouver, Canada, February 2024.


[Lo et al., 2021] Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray Buntine. Transformer over pre-


[Ma et al., 2023] Yongqiang Ma, Jiawei Liu, and Fan Yi. Is this abstract generated by AI? A research for the gap between AI-generated scientific text and human-written scientific text, 2023.


