Harnessing Dual-Role Context: A Noise-Suppressed Framework for Hybrid Human-AI Text Detection

Ying Wang, Yuling Liu, *Member, IEEE,* Ruyan Ding, Lingyun Xiang, *Member, IEEE,* Zhihua Xia, *Member, IEEE,* Xin Liao, *Senior Member, IEEE*

***Abstract*—Hybrid texts that combine human-written text (HWT) and AI-generated text (AIGT) content are becoming increasingly prevalent, creating a growing need for fine-grained authorship detection. Existing methods suffer from two major limitations. First, document-level classifiers cannot accurately localize AI-generated segments inside a document. Second, context-aware methods can help sentence-level detection but often introduce contextual noise when adjacent segments originate from different authors. To address these challenges, we propose a dual-role context framework with explicit noise suppression for fine-grained hybrid text detection. Specifically, we introduce incremental-perplexity features to capture local coherence variations across neighboring segments and develop a confidence-enhanced decision mechanism that adaptively aggregates multi-window contextual information, suppressing noisy signals while emphasizing reliable evidence. Furthermore, our unified framework jointly performs segment localization and authorship classification, avoiding the error propagation in traditional two-stage pipelines. Extensive experiments on multiple hybrid text benchmarks demonstrate that the proposed method consistently outperforms state-of-the art baselines across diverse hybrid ratios and domains. The code is available at: https://github.com/visionwy/Harnessing-Dual-Role-Context.**

***Index Terms*—AI text detection, coherence feature, confidence enhancement**

I. Introduction

THE rapid emergence of large language models (LLMs) has fundamentally reshaped automated text generation, producing text content with remarkable linguistic richness and stylistic fluency. Modern LLMs exhibit lexical, syntactic, and pragmatic features that closely mirror human writing, sometimes to the point of being indistinguishable [1]. In particular, some evaluations suggest that certain LLMs (e.g., those released by OpenAI, Google, and Anthropic) can effectively pass the Turing test, posing a significant challenge to traditional AI-generated text (AIGT) detection frameworks.

This work was supported in part by the National Natural Science Foundation of China under grant No. 62572176 and U23B2023, National Key Research and Development Program of China under grant No. 2024YFF0618800, and the Science and Technology Innovation Program of Hunan Province under grant 2025RC3166(Corresponding authors:Yuling Liu)

Ying Wang, Yuling Liu, Ruyan Ding, and Xin Liao are with the College of Cyber Science and Technology, Hunan University, Changsha 410082, China (e-mail: visionwang@hnu.edu.cn; yuling_liu@hnu.edu.cn; dingruyan@hnu.edu.cn; xinliao@hnu.edu.cn).

Lingyun Xiang is with the School of Computer Science and Technology, Changsha University of Science and Technology, Changsha 410076, China (e-mail: xiangly@csust.edu.cn).

Zhihua Xia is with the College of Cyber Security, Jinan University, Guangzhou, 510632, China(e-mail: xia_zhihua@163.com).

While LLMs have improved productivity in areas such as legal documentation, education, and journalism, their widespread deployment has also raised serious ethical and societal concerns. For example, [2] reported that nearly one-third of students had used ChatGPT to complete academic assignments. Meanwhile, the increasing accessibility of LLMs enables web bots to autonomously generate large volumes of persuasive content at negligible cost, facilitating public-opinion manipulation and the fabrication of social narratives. These developments underscore the urgent need for robust, fine-grained frameworks to detect AIGT.

Conventional AIGT detection methods primarily use a document-level binary classification paradigm, labeling an entire text as either human-written text (HWT) or AIGT. However, such coarse-grained approaches struggle in real-world settings, especially for increasingly common hybrid-origin texts that interweave human-written and AI-generated content. In practice, hybrid texts have become a dominant mode of content creation in journalism, education, and online platforms, making fine-grained detection essential for maintaining academic integrity, combating disinformation, and ensuring accountability in AI-assisted writing. This limitation motivates a key research question: How can we accurately localize AI-generated sentences in hybrid texts?

As illustrated in Fig.1(A), sentence-level detection is challenging because individual sentences are short, which limits discriminative lexical and semantic cues and reduces detection accuracy. Document-level detectors do not transfer directly to sentence-level tasks. Because they rely on document-level statistical and structural features, they are ill-suited to single sentences, which lack macro-level signals and sufficient context. To incorporate context, some frameworks adopt strategies such as the sliding-window method shown in Fig.1(B). This approach expands the analysis from an isolated sentence to a segment that includes the target sentence, using surrounding text to support classification and improving performance on short inputs.

However, contextual information plays a dual role in hybrid text detection. On the one hand, context can provide complementary evidence that strengthens the representation of a target sentence when neighboring segments share the same authorship. On the other hand, authorship transitions introduce heterogeneous context that may contaminate sentence representations and mislead the classifier. Despite its importance, this dual nature of context has received limited attention in existing studies. Most methods either ignore contextual cues or treat all surrounding text as equally informative, making them particularly vulnerable to errors around authorship boundaries.

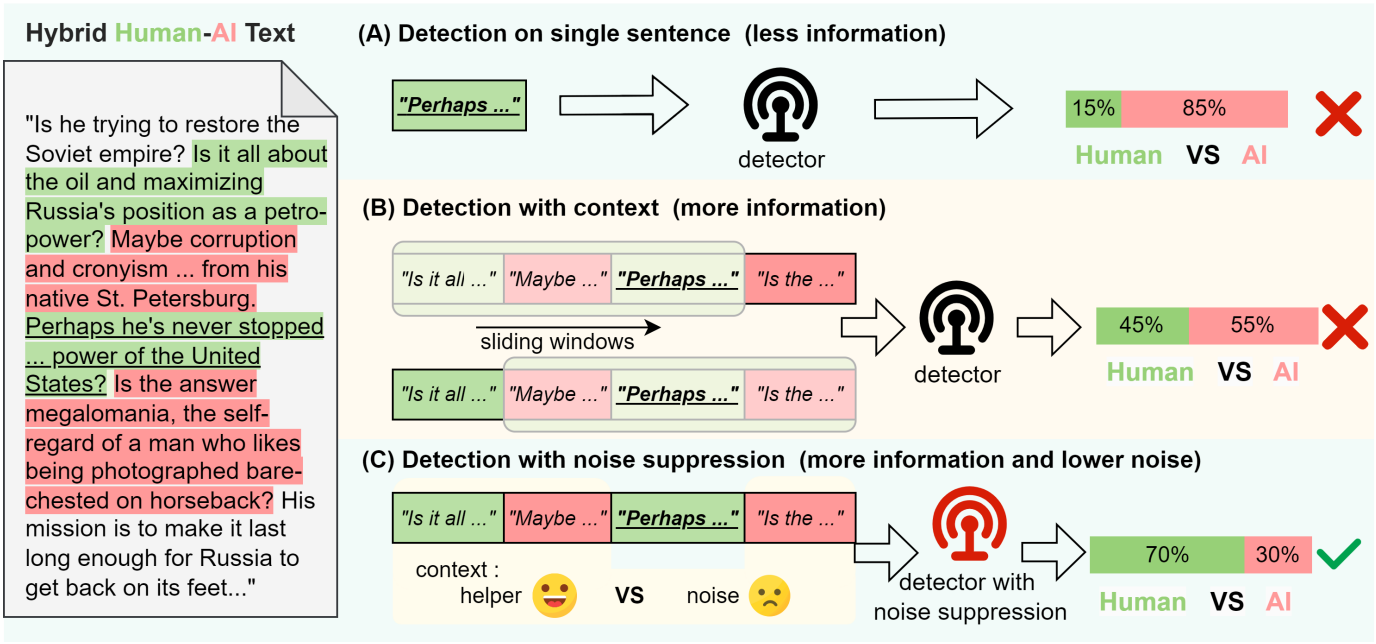


Fig. 1. Prior work in hybrid human-AI text detection.

This gap motivates a key question: How can contextual information be effectively leveraged for fine-grained hybrid text detection while suppressing noise from heterogeneous surrounding text?

As shown in Fig.1(C), we propose a dual-role context framework that treats contextual information as both a potential source of evidence and a source of noise. This framework is motivated by two fundamental characteristics of hybrid-origin texts. First, neighboring sentences with different authorship labels may introduce contextual label interference, causing contextual representations to deviate from the true characteristics of the target sentence. Second, transitions between human-written and AI-generated content often lead to local coherence disruption, reflecting the distributional divergence between the two writing styles. To capture these effects, we design incremental perplexity features that measure fine-grained coherence variations between a target sentence and its surrounding context. Based on these features, we further develop a confidence-enhanced decision mechanism that performs multi-window contextual fusion, selectively amplifying reliable evidence while suppressing conflicting contextual signals. By integrating segment localization and authorship classification into a unified end-to-end architecture, the proposed framework avoids the error accumulation inherent in conventional two-stage pipelines and achieves more robust fine-grained hybrid text detection.

The main contributions of this paper are summarized as follows:

- Dual-role context modeling. We identify the dual nature of contextual information in hybrid texts: neighboring segments can either provide supportive evidence or introduce misleading signals depending on authorship consistency. To characterize this phenomenon, we develop coherence-aware features based on incremental perplexity variations, enabling effective discrimination between beneficial and noisy contextual information.
- Noise-suppressed detection mechanism. We propose a confidence-enhanced verification mechanism that adaptively evaluates contextual reliability and integrates multi-window evidence. By reinforcing high-confidence predictions while suppressing conflicting contextual signals, the proposed mechanism enables a self-correcting and more robust detection process.
- Unified end-to-end framework. We design a unified framework that jointly performs AI-generated segment localization and authorship classification. Unlike conventional two-stage pipelines, the proposed framework mitigates error propagation and enables more effective interaction between the two tasks during training and inference.
- Comprehensive experimental evaluation. Extensive experiments on multiple benchmark datasets demonstrate that the proposed method consistently outperforms state-of-the-art approaches across diverse hybrid-content ratios and domains.

II. Related Work

The rapid development of AIGT detection technology is essential for safeguarding the security of LLM systems. While existing methods perform reliably in document-level binary classification, they face substantial challenges in hybrid text scenarios where human-written and AI-generated passages are interwoven.

Many hybrid text AIGT detection approaches use text segmentation to enable finer-grained analysis. Fine-grained detection based on segmentation generally follows two directions. The first direction is sentence-level sequence labeling. For instance, SeqXGPT [3] formulates the task as sequence tagging, leverages the log-probability vectors from white-box LLMs as

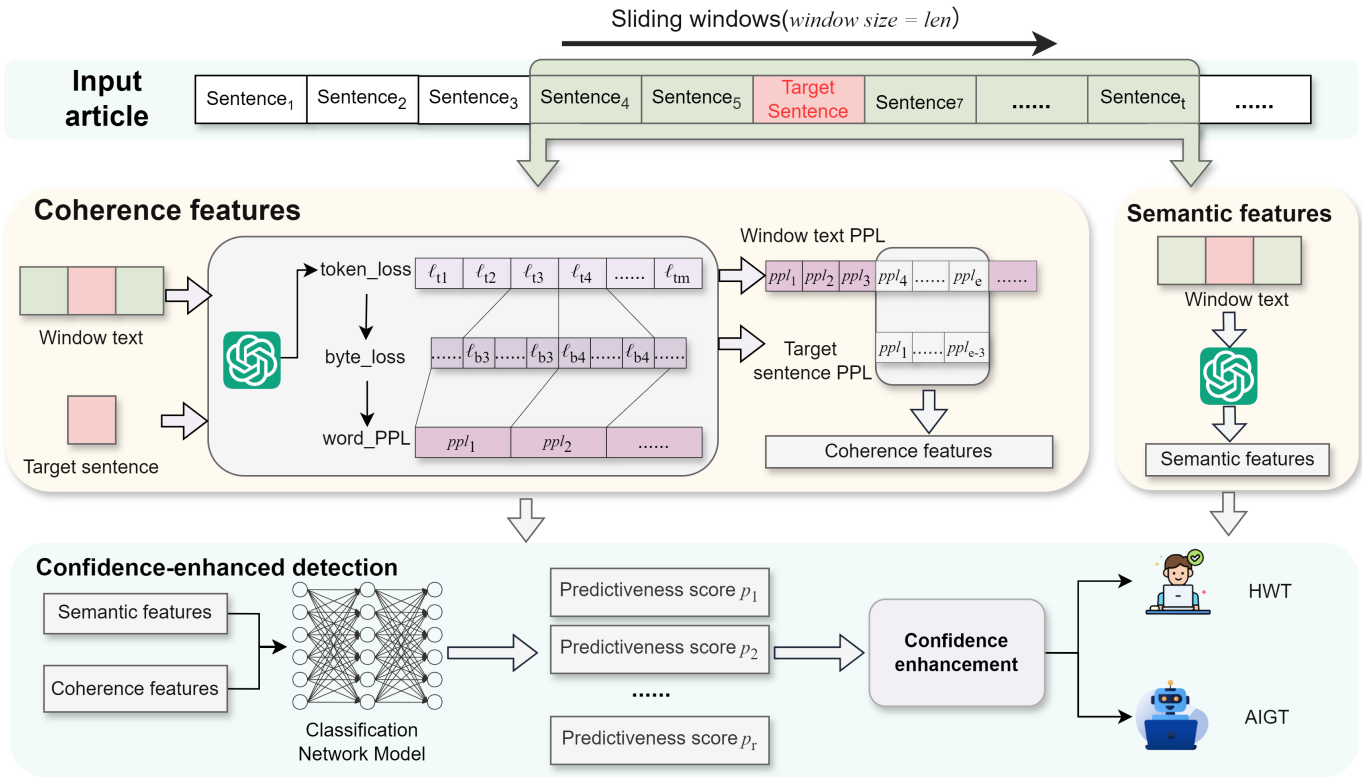


Fig. 2. Overview of noise-suppressed framework for hybrid human-AI text detection.

discriminative features, and ultimately delivers sentence-level predictions. Another sentence-level method [4] focuses on detecting text segments interpreted by LLMs, which splits text into sentences and designs a loss function based on semantic similarity to optimize the classifier, thereby identifying AI-generated portions. The approach in [5] segments both AIGT and HWT into sentences, computes similarity to construct features, and trains a classifier. The second direction operates at the paragraph or segment level. A multiscale positive-unlabeled framework introduced in [6] effectively enhances detection performance on short texts by incorporating a length-sensitive multi-scale positive-unlabeled loss function. The systematic comparison of segmentation-classifier combinations in [7] demonstrates that TriBERT [8] + DeBERTaV3 [9] achieves the most notable results. These studies indicate that fine-grained segmentation helps improve detection accuracy for hybrid texts, yet such methods often accumulate errors and underutilize contextual dependencies.

Given the limited discriminative information contained in individual sentences or fragments, some studies emphasize incorporating contextual information to boost detection performance. The core idea is to leverage text surrounding the target unit as auxiliary evidence. One intuitive technique is sliding window with aggregated prediction. The method proposed in [10] processes windows containing multiple sentences at once and aggregates multiple predictions of the same sentence across different windows via majority voting, thereby using context to stabilize sentence-level judgments. Another strategy involves generating contrastive samples from context. The SenDetEX framework [11] uses the preceding text of a target sentence to generate a semantically similar "regenerated sentence." By comparing embedding differences between the original and regenerated sentences, the method captures contextual coupling strength, thereby introducing rich contextual features. The methodology put forward by [12] designs a lightweight supervisor for long-text detection, whose core is a progressive supervision strategy aimed at enhancing the model's ability to discern complex contextual settings. Such methods confirm that context can offer valuable complementary signals, especially when stylistic cues are faint in short texts. However, in hybrid texts with frequent authorship switches, adjacent context may itself originate from different authorship. Undifferentiated use of such context may introduce noise and amplify the risk of misclassification.

Meanwhile, other studies have advanced hybrid text detection from varied perspectives. For example, evaluation benchmarks constructed in [13], [14] reveal the high false alarm rates and model biases of existing detectors when handling AI-augmented or hybrid content. A hierarchical detection framework proposed in [15] aims to differentiate levels of AI involvement. The generative source structure of text is modeled in [16] by building a hierarchical affinity tree to improve generalization. Additionally, the boundary detection task in SemEval-2024 Task 8 [17] has spurred research on word-level localization; top-performing solutions often combine strong encoders (e.g., DeBERTa) with sequence models such as conditional random fields [18]–[20].

In summary, current hybrid human-AI text detection methods have three main limitations:(1) two-stage pipelines decouple localization and classification, preventing joint optimization;

(2) context is often exploited without dynamically identifying and suppressing noisy components; and (3) the role of textual-coherence features in contextual fusion and noise suppression remains underexplored. To address these limitations, we propose a novel detection framework. Our core contribution lies in introducing the notion of dual-role context and to fuse it with textual-coherence features. Specifically, we design an incremental perplexity measure that converts coherence signals into dynamic features for suppressing context noise from heterogeneous authorship and enabling confidence-enhanced verification. Moreover, unlike conventional two-stage pipelines, our model jointly learns localization and classification within a single end-to-end framework.

III. Methods

Given a hybrid human-AI text $S = \{s_1, s_2, \cdots, s_n\}$ comprising n sentences, where each s_i may originate from either human or AI. Our framework aims to generate a probabilistic label sequence $Y = \{y_1, y_2, \cdots, y_n\}$, where $y_i \in [0, 1]$ quantifies the likelihood of s_i being AIGT. The overall architecture is shown in Fig. 2, which consists of three major modules: the semantic feature extraction module, the coherence feature extraction module and the confidence-enhanced detection module.

Our framework first processes hybrid text S through a dynamic sliding-window mechanism to capture local context. Each analysis window $T = \{t_1, \cdots, t_{len}\}$ represents a contiguous segment of length len drawn from S, where each t_i corresponds to a sentence in the original text.

A. Preliminaries and Notation

For clarity, the key symbols used in this section are defined in Table I.

TABLE I
List of key symbols used in Section III.

Symbol	Description
B_j	Byte sequence of the j-th token
c_i	Raw confidence score of the i-th prediction
D_k	Set of byte indices belonging to word h_k
E_{byte}	Predefined byte-encoding table used by the BBPE tokenizer
E_{byte}^{-1}	Inverse byte-encoding map, decoding a token into its byte sequence
f_{coh}	Coherence feature vector extracted from a window
$G(\cdot)$	Perplexity computed by the base language model
$G_{\text{word\_t}}$	Word-level perplexity sequence of the target sentence evaluated independently
$G_{\text{word\_T}}$	Word-level perplexity sequence of the full window text
G'_{word}	Perplexity sequence of the target sentence extracted from the window-level evaluation
I_{map}	Byte-to-word mapping table, recording which word each byte belongs to
ℓ_j	Token-level cross-entropy loss for predicting token x_{j+1}
$\bar{\ell}_k$	Average byte-level loss for word w_k
$\mathcal{L}_{byte}$	Byte-level cross-entropy loss sequence
len	Window length, i.e., number of sentences per window
m	Number of tokens in the tokenized window text T
n	Total number of sentences in a hybrid text
p_i	Prediction probability for a target sentence in the i-th window
P	Prediction sequence for a sentence accumulated across windows
ppl_k	Word-level perplexity of word h_k
r	Number of predictions accumulated for a sentence
S	A complete hybrid text consisting of n sentences
s_i	The i-th sentence in text S
t	Target sentence within a window
T	A contiguous window segment drawn from S
w_i	Normalized confidence weight for the i-th prediction
h_k	The k-th word in a text
y	Aggregated final prediction for a sentence
ΔG	Incremental perplexity, measuring the influence intensity of context on the target sentence

B. Semantic feature extraction

Semantic features capture high-level linguistic information arising from lexical, syntactic, and discourse-level interactions. In hybrid text detection, such representations are particularly valuable because human-written text (HWT) and AI-generated text (AIGT) often exhibit subtle semantic and distributional differences that are difficult to identify through surface-level features alone. To obtain robust semantic representations, we employ a pre-trained language model (PLM), which effectively encodes contextualized linguistic knowledge and has demonstrated strong performance across a wide range of downstream classification tasks.

For each text window T, semantic features are extracted from the embedding of the special token [CLS] produced by the PLM encoder. As a global representation learned through self-attention, the [CLS] embedding aggregates contextual information from all tokens within the window and provides a compact fixed-dimensional semantic representation.

C. Coherence feature extraction

Perplexity quantifies the predictive uncertainty of a language model on a given text sequence. Since token probabilities are conditioned on the surrounding context, perplexity is inherently context-dependent and reflects the degree of semantic compatibility between a sentence and its contextual environment. As a result, identical sentences may yield substantially different perplexity values when placed in different contexts. Such variations capture local coherence shifts and can therefore serve as indicators of authorship transitions in hybrid-origin texts. Based on this observation, we formulate this effect as follows:

$$\Delta G = G(t|T) - G(t|\varnothing) \tag{1}$$

Here $G(t|\varnothing)$ denotes the perplexity of the target sentence evaluated without contextual information, reflecting its independent linguistic representation, while $G(t|T)$denotes the perplexity of the same sentence conditioned on the surrounding window text, capturing the influence of context on its prediction probability. The difference ΔG thus quantifies the aggregate effect of context on the target sentence.

This incremental perplexity ΔG reflects two competing contextual mechanisms. On the one hand, semantic reinforcement

occurs when the context is coherent with the target sentence in terms of discourse continuity and authorship consistency, providing informative constraints that reduce the uncertainty of language model predictions. However, coherence disruption emerges at authorship boundaries, where heterogeneous contextual signals introduce distributional inconsistency within the window. Although the target sentence itself may remain fluent, the surrounding context may exhibit abrupt stylistic or topical transitions, thereby violating local coherence assumptions and degrading predictive stability. These opposing effects allow ΔG to serve as a discriminative indicator of authorship transitions in hybrid text.

The direct computation of token-level perplexity is sensitive to inconsistencies in tokenization schemes, which hinders the reliable cross-model alignment of likelihood estimates. To mitigate this issue, we adopt bytes as a unified representation unit and aggregate token-level losses to the word level, enabling consistent and tokenizer-agnostic perplexity computation. The overall procedure is summarized in Algorithm 1.

Algorithm 1 Word-Level Incremental Perplexity Computation.

Input: Window text T, target sentence t, base language model M with BBPE tokenizer

Output: Coherence feature f_{coh}

Step 1: Compute word-level perplexity for the whole window T

Tokenize T into $[x_1, x_2, \ldots, x_m]$.
Obtain token-level losses $\ell_1, \ldots, \ell_{m-1}$ from M.
Decode each token into bytes: $B_j \leftarrow \{E^{-1}_{\text{byte}}(c) \mid c \in x_j\}$.
Build byte-level loss sequence:
for $j \leftarrow 1$ **to** m **do**
  if $j = 1$ **then**
    Append $|B_1|$ zeros to $\mathcal{L}_{\text{byte}}$.
  else
    Append $|B_j|$ copies of ℓ_{j-1} to $\mathcal{L}_{\text{byte}}$.
  end if
end for
Compute word-level perplexity:
for each word h_k in T **do**
  $D_k \leftarrow \{d \mid I_{\text{map}}[d] = h_k\}$
  $\bar{\ell}_k \leftarrow \frac{1}{|D_k|} \sum_{d \in D_k} \mathcal{L}_{\text{byte}}[d]$
  $ppl_k \leftarrow \exp(\bar{\ell}_k)$
end for
Let $G_{\text{word\_T}}$ be the sequence of all ppl_k in order.

Step 2: Compute word-level perplexity for the target sentence t alone

Apply Steps 1's tokenization and loss extraction to t, obtaining its own token sequence and losses.
Perform the same byte remapping and word-level aggregation to obtain $G_{\text{word\_t}}$.

Step 3: Extract target's contextualized perplexity and compute difference

$G'_{\text{word}} \leftarrow$ truncate $G_{\text{word\_T}}$ to the word positions corresponding to t within T.
$\Delta G_{\text{word}} \leftarrow \dfrac{G_{\text{word\_t}} - G'_{\text{word}}}{G_{\text{word\_t}}}$
$f_{\text{coh}} \leftarrow \Delta G_{\text{word}}$

D. Confidence-enhanced detection

The semantic and coherence features extracted from the text are fed into a classification network for training. A fully connected layer is employed to capture complex nonlinear relationships among features and to transform them into categorical probability outputs.

Our detection framework employs a sliding-window traversal over the test datasets. For each test sample, the trained classifier processes the content within each window iteratively, yielding context-aware predictions at every step. This design enables each sentence to be evaluated from multiple contextual perspectives, yielding r prediction instances per sentence. For sentences located at the boundaries, the value of r is reduced. The resulting sequence of predictions reflects how sentence-level classification varies with the surrounding discourse:

$$P = \{p_i | \, i = 1, \cdots, r\} \tag{2}$$

where r denotes the number of times the sentence is predicted. Once the prediction sequence P is obtained for a sentence, we assign a confidence weight to each probability p_i. This weighting amplifies high-confidence predictions while down-weighting uncertain ones, enabling the model to prioritize reliable judgments and suppress noise. Specifically, for a sentence with prediction sequence $P = \{p_1, p_2, \ldots, p_r\}$, we first compute the raw confidence score for each p_i as the absolute distance from the neutral value 0.5:

$$c_i = |p_i - 0.5| \tag{3}$$

The normalized confidence weight w_i is then obtained by:

$$w_i = \frac{c_i}{\sum_{j=1}^{r} c_j} \tag{4}$$

To avoid division by zero, if all c_i are zero we assign equal weights $w_i = 1/r$.The final prediction y for the sentence is the weighted sum of the original probabilities:

$$y = \sum_{i=1}^{r} w_i \cdot p_i \tag{5}$$

Sentences located at the boundary positions of a document (i.e., the beginning or end) are typically covered by fewer overlapping sliding windows due to limited contextual availability. Our confidence-weighted aggregation mechanism is inherently robust to such boundary cases and does not rely on padding strategies or extrapolated contextual estimation. Specifically, when a sentence is associated with only a single prediction, its confidence weight is set to one, and the final output directly corresponds to that prediction. When multiple predictions are available, we perform confidence-based weighted aggregation to obtain a calibrated and stable decision. In all cases, the final prediction is computed solely based on the available contextual evidence.

IV. Experiments

A. Datasets and Metrics

- **Training and validation dataset:** Our experiments are conducted on the GoodNews dataset, a comprehensive corpus of New York Times articles from 2010 to 2018 [21]. To

simulate realistic hybrid text settings, we construct training and validation samples by blending original journalistic passages with text generated by GPT-2 [22]. In total, we construct 10,000 training window samples and 1,000 validation window samples, where each window sample contains a sequence of consecutive sentences whose length is determined by the sliding window configuration.

- **Test datasets:** To evaluate the detection performance and generalization of our method, we construct diverse test datasets. We use three hybrid text benchmarks: GoodNews , VisualNews (a pooled set of articles from four major news agencies) [23], and WikiText (Wikipedia-style passages) [24]. The AIGT is produced by varied models including GPT-2-1.5B, GPT-Neo-2.7B [25], GPT-J-6B [26], OPT-2.7B [27], GPT-NeoX-20B [28], LLaMA-3-8B [29], Qwen-2.5-7B [30]. Each test sample is a complete hybrid text containing 1-3 AI segments. The test set comprises 1,000 full texts from VisualNews and 60 full texts from WikiText. We note that in all our training and test datasets, the minimum text length is greater than the sliding window size (which is set to 3 throughout the experiments). Consequently, even the boundary sentences receive at least one prediction, and the influence of boundary handling on overall performance is limited.
- **Metrics:** We use Average Precision (AP) and Accuracy (Acc) to quantify the prediction accuracy of texts sampled from a specific LLM. Additionally, we use mean Average Precision (mAP) and mean Accuracy (mAcc) to measure the average performance of the model across test sets generated by different language models.

B. Baselines

- **OpenAI-Detector** [31] is a model developed by OpenAI to identify text generated by the GPT-2 model. It is based on a fine-tuned RoBERTa model, trained by using the outputs from the 1.5B-parameter GPT-2 model. This detector can effectively predict whether a given text was generated by GPT-2.
- **RoBERTa-MPU** [6] is a short text detection framework that incorporates length-sensitive loss, thereby enhancing the performance of short text detection while maintaining the accuracy of long text detection.
- **Easy2Hard** [12] is the method that enhances machine generated text detection by leveraging an easy supervisor (trained on longer texts) to guide a more challenging target detector. Furthermore, adhering to the parameters and settings specified in their paper, we carried out experiments on the RADAR-E model.
- **Fast-DetectGPT** [32] is a zero-shot detection method that leverages conditional probability curvature. It replaces the perturbation step in DetectGPT [33] with a more efficient sampling step, achieving both higher efficiency and greater precision in detection results. While primarily designed for long-text detection, we have adapted it to the sentence-level in our experiments. To further enhance its performance on short texts and to provide a more context-aware baseline, we have integrated a sliding window strategy into Fast-DetectGPT, creating a variant called **Fast-DetectGPT-SW**. Specifically, for each target sentence, we extract multiple overlapping context windows using a fixed step size. The final prediction for the sentence is obtained by averaging the detection scores across all windows in which it appears. This modification aims to validate the positive impact of incorporating contextual information on short-text detection under a consistent multi-window aggregation scheme.
- **Binoculars** [34] is a zero-shot detection methodology that performs detection by leveraging contrastive scores derived from two pre-trained models. Similarly, we have incorporated the sliding window technique called **Binoculars-SW** in our experiments to validate the impact of context.
- **AdaLoc** [10] is the method for performing hybrid human-AI text detection, as introduced earlier. Furthermore, adhering to the parameters and settings specified in their paper, we carried out experiments on the RoBERTa-large model.

C. Experimental comparison with baselines

As shown in Table II, we trained and evaluated our method on GoodNews to validate accuracy and further tested generalization on the cross-domain VisualNews and WikiText datasets. To systematically investigate the effect of AIGT proportion, we conducted experiments with texts containing 1 to 3 AI-generated segments (Seg-1 to Seg-3). For fair comparison, all methods used an identical sliding window configuration (window length=3, step=1), which was supported by AdaLoc and verified as optimal in our ablation studies (Section IV-F).

We draw several conclusions from the table. First, sliding windows significantly improve baseline detection performance. Specifically, on GoodNews, Fast-DetectGPT-SW increases mean AP by 13.93% and mean Acc by 4.01% over Fast-DetectGPT, while Binoculars-SW improves mean AP by 14.00% and mean Acc by 4.90% over Binoculars. These gains confirm that contextual information benefits short-text detection. However, this advantage diminishes as the number of AI segments increases from 1 to 3 across all datasets, and on WikiText Seg-3, Fast-DetectGPT-SW even underperforms its non-window counterpart. This degradation occurs because frequent authorship switches introduce context from different origins, causing label interference and coherence disruption that can turn helpful cues into noise.

Second, although our method also uses a sliding window, it suppresses noise more effectively than other sliding-window baselines. For instance, Fast-DetectGPT-SW obtains slightly higher Acc on VisualNews Seg-1 and WikiText Seg-1/2, highlighting the advantage of the pure sliding window method in scenarios with strong local features. However, its performance drops sharply on Seg-3: from 86.62% to 81.08% on GoodNews (a 5.54% drop) and from 89.88% to 85.13% on VisualNews (a 4.75% drop). In contrast, our method maintains segment differences within 0.9% on GoodNews and 0.47% on VisualNews. This contrast indicates that expanding context without active noise suppression can introduce instability,

TABLE II
DETECTION RESULTS ACROSS MULTIPLE DOMAINS AND WITHIN INDIVIDUAL DOMAINS

Dataset	Method	Seg-1		Seg-2		Seg-3		Mean	
		AP	Acc	AP	Acc	AP	Acc	AP	Acc
GoodNews (in-domain)	OpenAI-Detector [31]	30.24	60.50	32.20	60.81	33.33	61.06	31.92	60.79
	RoBERTa-MPU [6]	67.31	83.26	68.71	82.57	69.38	82.28	68.47	82.70
	Easy2Hard [12]	55.33	66.89	58.43	67.63	59.44	67.69	57.73	67.40
	Binoculars [34]	36.05	75.32	38.07	74.41	39.27	74.10	37.80	74.61
	Binoculars-SW	56.38	81.59	51.30	79.15	47.64	77.78	51.78	79.51
	Fast-DetectGPT [32]	49.14	80.11	51.88	79.31	53.38	79.11	51.47	79.51
	Fast-DetectGPT-SW	73.10	86.62	63.99	82.85	59.09	81.08	65.39	83.52
	AdaLoc [10]	86.53	90.83	85.34	89.39	84.80	88.85	85.55	89.69
	Ours	**89.49**	**92.87**	**89.56**	**92.20**	**89.92**	**91.97**	**89.65**	**92.35**
VisualNews (cross-domain)	OpenAI-Detector [31]	24.45	63.40	25.74	63.56	25.68	63.68	25.29	63.55
	RoBERTa-MPU [6]	57.19	83.73	58.06	83.42	58.77	83.77	58.00	83.64
	Easy2Hard [12]	43.29	65.65	45.17	66.02	44.94	66.06	44.47	65.90
	Binoculars [34]	31.72	79.69	32.81	79.02	32.05	79.13	32.19	79.28
	Binoculars-SW	54.61	85.81	45.95	84.15	36.12	83.48	45.56	84.48
	Fast-DetectGPT [32]	48.71	85.22	49.54	84.57	48.75	84.64	49.00	84.81
	Fast-DetectGPT-SW	74.35	**89.88**	61.72	86.62	49.07	85.13	61.71	87.21
	AdaLoc [10]	74.23	86.37	73.14	85.48	71.12	85.08	72.83	85.64
	Ours	**77.45**	88.45	**77.80**	**88.21**	**77.75**	**87.99**	**77.67**	**88.21**
WikiText (cross-domain)	OpenAI-Detector [31]	22.41	65.72	24.95	65.87	24.32	66.23	23.89	65.94
	RoBERTa-MPU [6]	56.84	86.48	52.55	84.73	54.89	85.21	54.76	85.47
	Easy2Hard [12]	44.62	68.96	47.25	68.30	48.29	67.98	46.72	68.41
	Binoculars [34]	31.09	81.91	30.40	81.23	33.21	81.63	31.56	81.59
	Binoculars-SW	48.01	88.18	46.68	86.43	33.21	85.34	42.63	86.65
	Fast-DetectGPT [32]	50.85	88.31	50.40	86.22	49.16	86.23	50.14	86.92
	Fast-DetectGPT-SW	74.89	**91.69**	68.77	**89.72**	45.77	86.18	63.15	**89.20**
	AdaLoc [10]	71.91	85.81	69.99	84.73	71.42	84.94	71.10	85.16
	Ours	**77.29**	89.53	**76.35**	88.40	**74.87**	**87.86**	**76.17**	88.60

whereas our approach improves both accuracy and stability by explicitly modeling coherence disruption.

Third, our method achieves optimal or near-optimal performance across all three datasets. On GoodNews, ours achieves a mean AP that is 4.10% higher than AdaLoc (the second-best method) and a mean Acc that is 2.66% higher. In addition, the segment-wise variation in AP and Acc is small (less than 0.5% and 0.9%, respectively), demonstrating strong stability. On the cross-domain VisualNews and WikiText datasets, our method achieves substantially higher mean AP than other baselines, further supporting its robustness across diverse text types. This stability stems from our coherence feature, which quantifies semantic inconsistency via perplexity differences, and confidence-weighted fusion, which down-weights unreliable window-level predictions.

These findings are consistent with prior observations that sliding-window aggregation methods degrade significantly when authorship transitions occur frequently within a document. For instance, AdaLoc relies on majority voting over overlapping windows without explicit modeling of contextual noise, which limits its ability to effectively suppress interference from heterogeneous surrounding contexts. Beyond sliding-window approaches, Easy2Hard attempts to enlarge distributional separability by concatenating multiple texts to amplify inter-class gaps. However, its AP and accuracy remain relatively low under our imbalanced evaluation setting, indicating limited robustness in realistic hybrid scenarios. In contrast, our framework explicitly models coherence disruption and incorporates uncertainty-calibrated prediction through confidence-aware aggregation, resulting in consistently robust performance even under data imbalance. Overall, the experimental results not only validate the benefit of contextual information but also highlight the importance of explicitly modeling and mitigating contextual noise, which has been largely underexplored in existing hybrid text detection literature.

D. Ablation experiment on GoodNews

Table III and Table IV summarize the ablation results in the GoodNews dataset, where 'Ours-f_{coh}' denotes our method without the coherence feature, 'Ours-*confi*' denotes the variant without confidence enhancement, and "Ours-*both*" removes both modules. The model is trained exclusively on GPT-2-1.5B-generated text, and evaluated under both intra-generator (GPT-2) and cross-generator settings, including GPT-Neo, OPT, LLaMa and Qwen. To further assess robustness, we conducted experiments using two strong backbone encoders, RoBERTa-large and DeBERTa-V3-large, representing different levels of contextual encoding capacity.

The ablation results yield several key findings. Both modules consistently improve performance, and their integration achieves the best overall results. Using RoBERTa as the backbone, the full model (Ours) achieves 81.15% mAP and 88.34% mAcc, surpassing the Ours-*both* by 1.52% and 1.01%, respectively. The gains remain consistent across different

TABLE III
AP RESULTS OF THE SAME GENERATOR AND ACROSS GENERATORS

Method	GPT-2	GPT-Neo	OPT	LLaMa	Qwen	mAP
(A) Experiment on RoBERTa						
Ours-*both*	85.55	84.72	79.88	74.89	73.10	79.63
Ours-f_{coh}	85.66	84.81	80.03	74.92	73.49	79.78
Ours-*confi*	86.59	85.68	80.59	76.08	75.09	80.80
Ours	**86.80**	**85.86**	**80.83**	**77.16**	**75.11**	**81.15**
(B) Experiment on DeBERTa						
Ours-*both*	89.38	88.86	82.28	80.15	79.01	83.94
Ours-f_{coh}	89.38	88.90	82.31	80.43	79.15	84.03
Ours-*confi*	**89.69**	89.58	82.81	80.55	79.82	84.49
Ours	89.65	**89.60**	**82.84**	**81.25**	**80.09**	**84.69**

TABLE IV
ACC RESULTS OF THE SAME GENERATOR AND ACROSS GENERATORS

Method	GPT-2	GPT-Neo	OPT	LLaMa	Qwen	mAcc
(A) Experiment on RoBERTa						
Ours-*both*	89.69	89.70	88.30	84.38	84.61	87.33
Ours-f_{coh}	89.86	89.79	88.43	84.40	84.69	87.43
Ours-*confi*	90.95	90.79	89.13	84.68	84.71	88.05
Ours	**91.13**	**90.95**	**89.32**	**85.47**	**84.82**	**88.34**
(B) Experiment on DeBERTa						
Ours-*both*	91.98	91.91	89.52	86.71	86.07	89.24
Ours-f_{coh}	92.04	91.97	89.56	87.30	86.25	89.42
Ours-*confi*	92.21	92.55	89.96	87.33	86.33	89.68
Ours	**92.35**	**92.57**	**90.00**	**87.42**	**86.50**	**89.77**

backbones and all test generators, demonstrating strong robustness and generalization. These observations suggest that the two modules address complementary aspects of the problem. Specifically, the coherence feature captures local semantic irregularities indicative of authorship transitions, whereas the confidence-based aggregation reduces prediction variance by suppressing noisy contextual signals. Their combination enables more stable and discriminative modeling of hybrid text.

Second, backbone architecture plays a significant role in performance, with DeBERTa consistently outperforming RoBERTa across all settings. Under the full model, DeBERTa achieves 84.69% mAP and 89.77% mAcc, excedding RoBERTa by 3.54% and 1.43% points, respectively. This trend is stable across all ablation variants, suggesting that the improvement is primarily driven by the representational capacity of the backbone rather than specific module interactions. The superior performance of DeBERTa can be attributed to its disentangled attention mechanism and improved pre-training design, which facilitate more precise modeling of subtle lexical and syntactic differences between human-written and AI-generated text.

Third, the coherence feature plays a more critical role than the confidence enhancement module in overall performance. The removal of the coherence component (Ours-f_{coh}) leads to a significantly larger performance drop compared to removing the confidence module (Ours-*confi*), indicating that it is the primary contributor to the model's effectiveness. This finding underscores the importance of explicitly modeling semantic inconsistency at authorship transition boundaries for fine-grained hybrid text detection.

E. Cross-domain and cross-generator evaluation with varying segments

To comprehensively evaluate the robustness and generalization capability of our method, we consider three challenging factors, including varying numbers of AIGT segments, different text generators, and unseen domains. All models are trained solely on the GoodNews-GPT2 dataset and evaluated under two complementary scenarios. The first scenario focuses on in-domain robustness, using GoodNews test sets containing one to three AIGT segments generated by five different language models. The second scenario evaluates cross-domain generalization on the unseen VisualNews and WikiText datasets, providing a stringent assessment of the model's resilience to domain shift.

In the following figures, we use consistent markers to denote different variants of our approach for clarity: 'A' represents the "Ours-*both*" method , 'B' denotes the "Ours-*confi*" method, 'C' stands for the "Ours-f_{coh}" method, and 'D' corresponds to our full proposed method. The average results across generators for each segment count are shown in Fig. 3 (detailed per-generator results are provided in the Appendix B). The domain generalization results on VisualNews and WikiText are presented in Fig. 4 and Fig. 5, respectively.

Our analysis leads to the following key observations. First, the choice of base encoder significantly impacts stability. On the GoodNews dataset, when using DeBERTa as the base model, performance remains highly stable as the number of AIGT segments increases from 1 to 3, with fluctuations within 1%. In contrast, the RoBERTa-large model exhibits a noticeable decline under the same condition, with mAP and mAcc dropping by an average of 3.98% and 2.43%, respectively.Second, both proposed components yield measurable gains, though the perplexity-based coherence feature contributes more substantially to the overall improvement than the confidence-enhancement mechanism. Finally, our method demonstrates strong generalization ability. On the VisualNews dataset, it achieves an average improvement of 1.01% AP and 2.78% Acc with RoBERTa, and 3.21% AP and 0.67% Acc with DeBERTa. On the WikiText dataset, the gains are even more substantial, with average improvements of 5.62% AP and 5.39% Acc under RoBERTa, and 2.80% AP and 2.20% Acc under DeBERTa. These results confirm that our framework effectively handles varying generator outputs and generalizes well to unseen domains.

F. Sensitivity analysis of window parameters

To systematically evaluate the effect of sliding-window parameters on detection performance, we conduct a sensitivity analysis over window sizes from 2 to 6 and step sizes from 1 to 6. For each parameter setting, the model is evaluated independently on test sets containing one, two, and three AI-generated segments. The AP and Acc results shown in Fig. 6. correspond to the average performance across these three scenarios, thereby providing a robust and comprehensive evaluation of parameter sensitivity under varying levels of hybrid-content complexity.

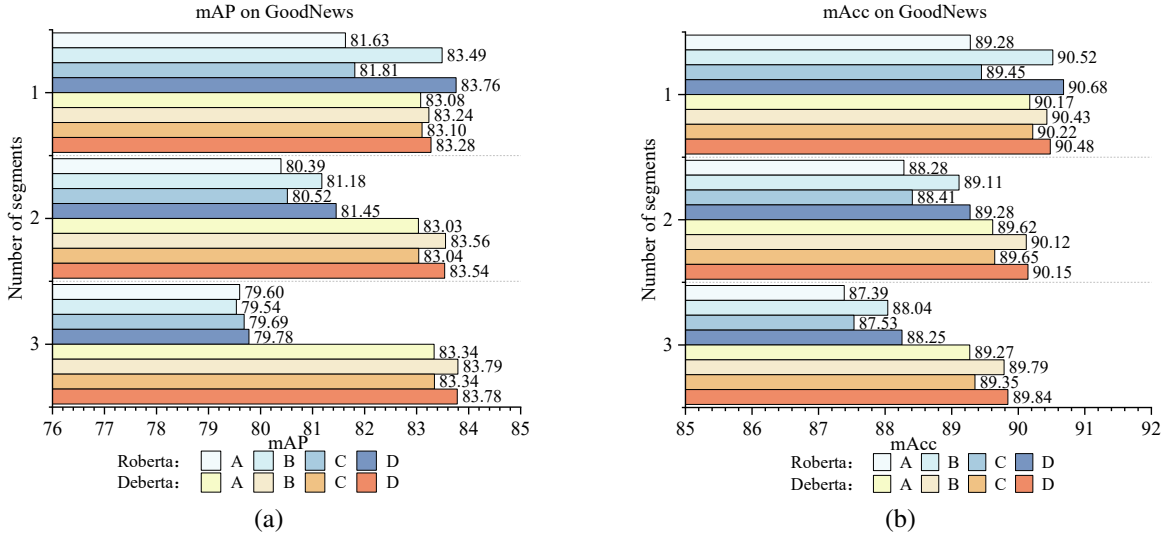


Fig. 3. Cross-generator average detection performance on GoodNews with varying numbers of AIGT segments: (a) mAP and (b) mAcc.

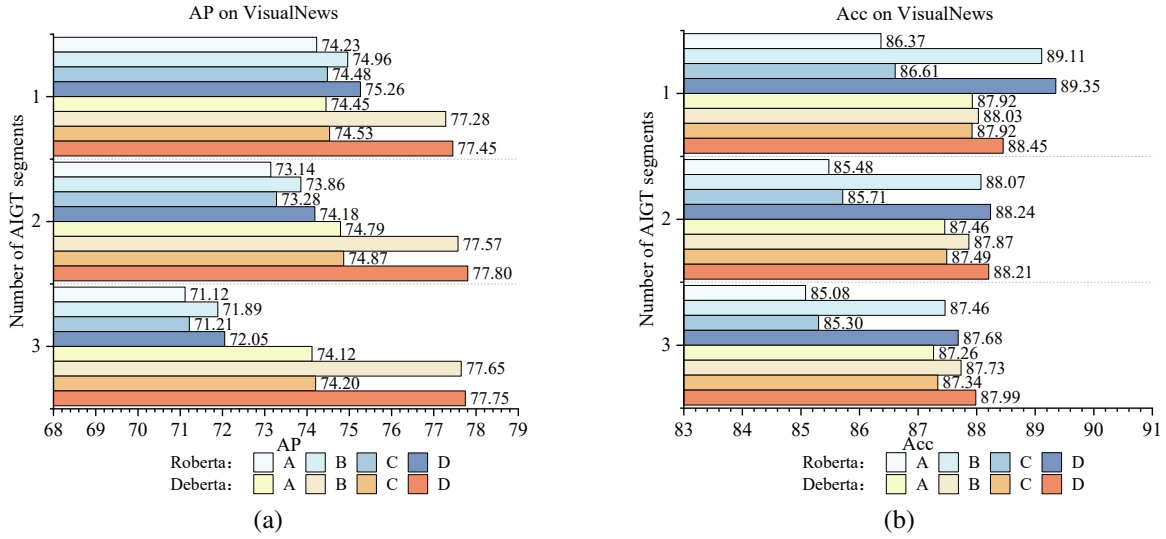


Fig. 4. Cross-domain evaluation results on VisualNews with varying numbers of AIGT segments: (a) AP and (b) Acc.

The experimental results reveal a clear context–noise trade-off. When the window size is small, the available contextual information is limited, preventing the model from fully capturing semantic dependencies and authorship-related patterns. Consequently, performance remains relatively low (e.g., AP=74.89%, Acc=86.15% at window size=2, step=1). Increasing the window size to 3 significantly improves performance, indicating that moderate contextual expansion provides informative evidence for identifying authorship transitions. However, enlarging the window beyond this point produces diminishing returns. While additional context may introduce useful information, it also increases the probability of incorporating semantically unrelated or differently authored sentences. As a result, the benefit of contextual enrichment is gradually offset by the accumulation of contextual noise. This observation further validates the motivation behind our dual-role context framework and suggests that a moderate window size offers the best balance between contextual utility and noise suppression.

As for the step size, setting it to 1 allows each sentence to be evaluated in multiple overlapping contexts, which improves judgment accuracy and robustness. This setting yields consistently strong results across different window sizes. In contrast, larger steps (≥ 2) reduce contextual overlap, often omitting relevant transitional signals and resulting in noticeable declines in both AP and Acc.

Overall, the combination of a window size of 3 and a step size of 1 yields the most favorable performance across the majority of experimental settings. This configuration provides sufficient contextual information to enhance discriminability while ensuring that each sentence is examined from multiple contextual perspectives through fine-grained sliding.We therefore adopt these as the default parameters in our framework. This sensitivity study confirms the importance of parameter tuning and highlights the effectiveness of the

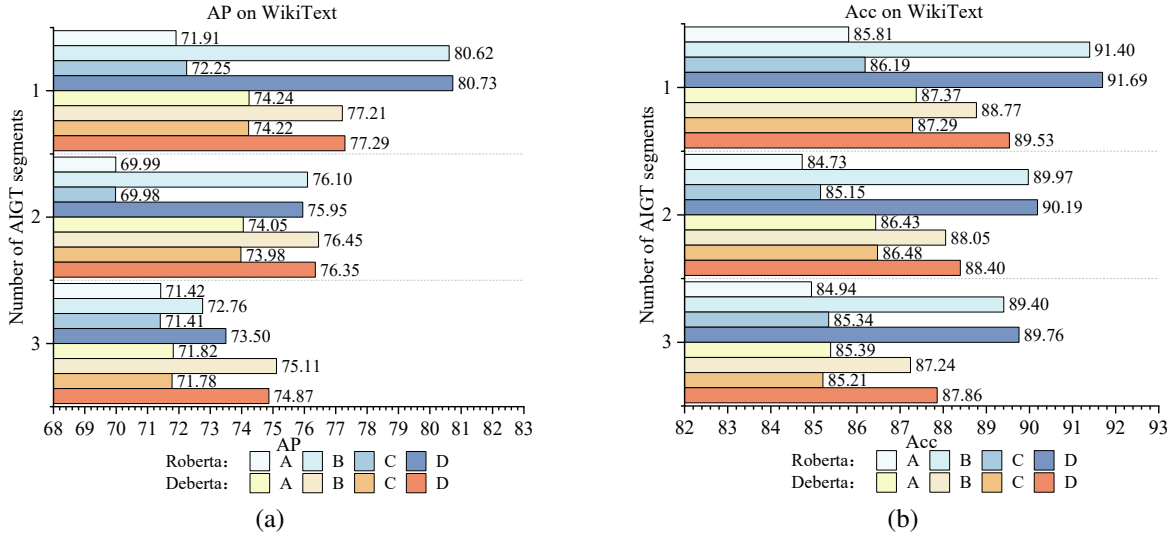


Fig. 5. Cross-domain evaluation results on WikiText with varying numbers of AIGT segments: (a) AP and (b) Acc.

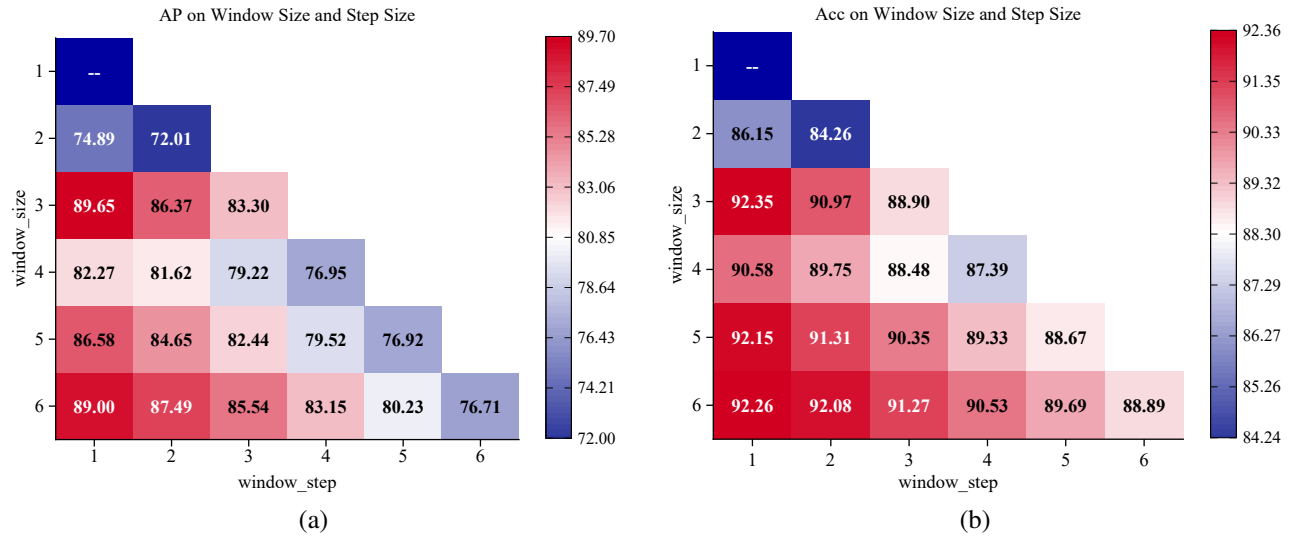


Fig. 6. Ablation experiment results about window size and step size: (a) AP and (b) Acc.

proposed noise-suppression and context-fusion strategy.

G. Experiment about target sentence

The experimental results reveal two key findings. First, DeBERTa serves as a more effective backbone, consistently delivering superior and more stable performance across diverse evaluation scenarios. Second, the proposed perplexity-based coherence feature provides a significant performance boost, demonstrating the value of explicitly capturing semantic consistency and authorship-transition signals. These observations motivate a closer examination of how the backbone architecture and coherence feature contribute to the overall effectiveness of the proposed framework.

Building on these findings, we explored the impact of designating multiple target sentences within the DeBERTa model. Specifically, we selected both the middle and the last sentences as target sentences in the same experiment. This setup allowed us to evaluate both the coherence of the first half of the text and the overall coherence. The results are detailed in Table V. In table, "Ours-f_{coh}" denotes our method without coherence feature, "Ours+L" denotes the approach of utilizing the last sentence within the window text as the target sentence for extracting coherence features, and "Ours+M&L" denotes the approach of utilizing both the middle and the last sentence.

From Table V, we observe that fusing multiple target positions (Ours+M&L) improves both mAP and mAcc compared to the single position strategy (Ours+L) and the variant without coherence features (Ours-f_{coh}). For instance, Ours+M&L achieves the highest mAP of 83.20% and mAcc of 89.56%, outperforming Ours+L (82.71% mAP, 89.24% mAcc) and Ours-f_{coh}(82.13% mAP, 88.74% mAcc). This indicates that incorporating both middle and last sentences provides richer contextual information: the middle sentence captures local coherence patterns within the window, while the last sentence

TABLE V
RESULTS REGARDING THE TARGET SENTENCES, WHICH WERE OBTAINED ON THE TEST SETS GENERATED BY DIFFERENT LLMS

Method	GPT-2		GPT-Neo		OPT		GPT-J		GPT-NeoX		LLaMa		Qwen		mAP	mAcc
	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc		
Ours-f_{coh}	89.38	91.98	88.86	91.91	82.28	89.52	80.87	88.22	74.35	86.80	80.15	86.71	79.01	86.07	82.13	88.74
Ours+L	89.65	92.35	89.60	92.57	82.84	90.00	81.16	88.76	74.41	87.11	81.25	87.42	80.09	86.50	82.71	89.24
Ours+M&L	**90.14**	**92.80**	**89.97**	**92.84**	**83.33**	**90.35**	**82.29**	**89.29**	**75.03**	**87.43**	**81.52**	**87.48**	**80.11**	**86.75**	**83.20**	**89.56**

reflects overall discourse coherence. Their combination yields more robust detection, particularly for high-quality AI texts where subtle inconsistencies require multiple positional views.

V. CONCLUSION

This paper proposes a dual-role context framework for fine-grained hybrid text detection, motivated by the observation that contextual information can function either as supportive evidence or as a source of interference depending on authorship consistency. To capture this duality, we introduce an incremental perplexity feature that characterizes local coherence variations and a confidence-enhanced fusion mechanism that selectively reinforces reliable contextual evidence while suppressing noisy signals. The resulting unified framework jointly performs segment localization and authorship classification, mitigating the error propagation inherent in conventional two-stage approaches. Experimental results demonstrate that the proposed method achieves consistent improvements across varying hybrid-content ratios, multiple text generators, and unseen domains, confirming its effectiveness, robustness, and generalization ability. Future research will focus on extending the framework to multilingual and multi-domain environments, improving robustness against adversarial and distribution-shift attacks, and developing more adaptive mechanisms for context-aware authorship analysis.

APPENDIX A
A STEP-BY-STEP EXAMPLE OF WORD-LEVEL PERPLEXITY CALCULATION

In this section, we provide a step-by-step illustrative example of the byte-level perplexity calculation used in our coherence feature extraction in Fig. 7, serving as a clear and reproducible reference for implementing this core methodological component.

APPENDIX B
DETAILED RESULTS PER GENERATOR

In this section, we present in detail the experimental results for each generator's text under varying segment counts. The specific data can be found in Fig. 8 to Fig. 12.

REFERENCES

[1] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, Red Hook, NY, USA, May 2020, pp. 1877–1901.

[2] S. Pudasaini, L. Miralles-Pechuán, D. Lillis, and M. Llorens Salvador, "Survey on AI-generated plagiarism detection: The impact of large language models on academic integrity," *Journal of Academic Ethics*, vol. 23, no. 3, pp. 1–34, Nov. 2024.

[3] P. Wang, L. Li, K. Ren, B. Jiang, D. Zhang, and X. Qiu, "SeqXGPT: Sentence-level AI-generated text detection," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1144–1156.

[4] Y. Li, Z. Wang, L. Cui, W. Bi, S. Shi, and Y. Zhang, "Spotting AI's touch: Identifying LLM-paraphrased spans in text," in *Findings of the Association for Computational Linguistics*. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 7088–7107.

[5] A. Quidwai, C. Li, and P. Dube, "Beyond black box AI generated plagiarism detection: From sentence to document level," in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 727–735.

[6] Y. Tian *et al.*, "Multiscale positive-unlabeled detection of ai-generated texts," in *International Conference on Learning Representations*, vol. 2024, Vienna, Austria, 2024, pp. 34 849–34 866.

[7] Z. Zeng *et al.*, "Detecting AI-generated sentences in human-AI collaborative hybrid texts: Challenges, strategies, and insights," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. Jeju Island, South Korea: International Joint Conferences on Artificial Intelligence Organization, Aug. 2024, pp. 7545–7553.

[8] Z. Zeng, L. Sha, Y. Li, K. Yang, D. Gašević, and G. Chen, "Towards automatic boundary detection for human-ai collaborative hybrid essay in education," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 20, pp. 22 502–22 510, Mar. 2024.

[9] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing," in *the Eleventh International Conference on Learning Representations*, Kigali Rwanda, May 2023.

[10] Z. Zhang, W. Qin, and B. Plummer, "Machine-generated text localization," in *Findings of the Association for Computational Linguistics*. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 8357–8371.

[11] L. Jiang, D. Wu, and X. Zheng, "SenDetEX: Sentence-level AI-generated text detection for human-AI hybrid content via style and context fusion," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 5287–5302.

[12] C. Wu, Y.-m. Cheung, B. Han, and D. Lian, "Advancing machine-generated text detection from an easy to hard supervision perspective," in *Advances in Neural Information Processing Systems*, vol. 38. Curran Associates, Inc., Dec. 2025, pp. 150 210–150 258.

[13] Q. Zhang *et al.*, "LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected?" in *Findings of the Association for Computational Linguistics*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 409–436.

[14] S. Saha and S. Feizi, "Almost AI, almost human: The challenge of detecting AI-polished writing," in *Findings of the Association for Computational Linguistics*. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 25 414–25 431.

[15] G. Bao, L. Rong, Y. Zhao, Q. Zhou, and Y. Zhang, "Decoupling content and expression: Two-dimensional detection of AI-generated text," 2025, *arXiv:2503.00258* .

[16] Y. He, S. Zhang, Y. Cao, L. Ma, and P. Luo, "DETree: DEtecting human-AI collaborative texts via tree-structured hierarchical representation learning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, San Diego, USA, Dec. 2026.

[17] Y. Wang *et al.*, "SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection," in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 2057–2079.

[18] R. M. R. Kadiyala, "RKadiyala at SemEval-2024 task 8: Black-box word-level text boundary detection in partially machine-generated texts," in

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024). Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 511–519.

[19] A. Voznyuk and V. Konovalov, "DeepPavlov at SemEval-2024 task 8: Leveraging transfer learning for detecting boundaries of machine-generated texts," in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 1821–1829.

[20] Z. Guo *et al.*, "USTC-BUPT at SemEval-2024 task 8: Enhancing machine-generated text detection via domain adversarial neural networks and LLM embeddings," in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 1511–1522.

[21] A. F. Biten, L. Gomez, M. Rusiñol, and D. Karatzas, "Good News, Everyone! Context Driven Entity-Aware Captioning for News Images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, Jun 2019, pp. 12 458–12 467.

[22] A. Radford *et al.* (2019) Language models are unsupervised multitask learners. [Online]. Available: https://openai.com/blog/better-language-models/

[23] F. Liu, Y. Wang, T. Wang, and V. Ordonez, "Visual News: Benchmark and challenges in news image captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6761–6771.

[24] M. Stephen, X. Caiming, B. James, S. Richard *et al.*, "Pointer sentinel mixture models," in *the 5th International Conference on Learning Representations*, Toulon, France, Apr. 2017, pp. 24–26.

[25] L. Gao *et al.*, "The pile: An 800gb dataset of diverse text for language modeling," 2020, *arXiv:2101.00027* .

[26] B. Wang and A. Komatsuzaki. (2021) GPT-J-6B: A 6 Billion parameter autoregressive language model. [Online]. Available: https://github.com/kingoflolz/mesh-transformer-jax

[27] S. Zhang *et al.*, "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068* .

[28] S. Black *et al.*, "GPT-NeoX-20B: An open-source autoregressive language model," in *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 95–136.

[29] A. Grattafiori *et al.*, "The llama 3 herd of models," 2024, *arXiv:2407.21783* .

[30] Qwen *et al.*, "Qwen2.5 technical report," 2025, *arXiv:2412.15115* .

[31] I. Solaiman *et al.*, "Release strategies and the social impacts of language models," 2019, *arXiv:1908.09203* .

[32] G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang, "Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature," in *The Twelfth International Conference on Learning Representations*, Vienna Austria, May 2024.

[33] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, Hawaii, USA, Jul. 2023, pp. 24 950–24 962.

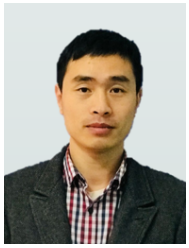
[34] A. Hans *et al.*, "Spotting LLMs with binoculars: zero-shot detection of machine-generated text," in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, Vienna, Austria, Jul. 2024, pp. 17 519–17 537.

Ying Wang received the B.S. degree in information security from Hunan University, Changsha, China, in 2023. She is currently working toward the M.E. degree in computer science and technology at Hunan University, Changsha, China. Her research interests include artificial intelligence security, machine learning, and natural language processing.

Yuling Liu (Member, IEEE) received her B.E. degree in computer science and technology, in 2003, and the Ph.D. degree in computer application, in 2008, Hunan University, Changsha, China. She is currently a professor with the College of Cyber Science and Technology of Hunan University, Changsha, China. Her research interests include AI security, information security, natural language processing, multimedia steganography.

Ruyan Ding received the B.S. degree in Data Science and Big Data Technology from Hainan University, Haikou, China, in 2024. She is currently working toward the M.E. degree in Cyberspace Security at Hunan University, Changsha, China. Her research interests include artificial intelligence security, cognitive security, and natural language processing.

Lingyun Xiang (Member, IEEE) received her B.E. degree in computer science and technology, in 2005, and the Ph.D. degree in computer application, in 2011, Hunan University, Changsha, China. She is currently a Professor with the School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, China. Her research interests include information security, steganography, and steganalysis, natural language processing, machine learning, pattern recognition.



Zhihua Xia (Member, IEEE) received his Ph.D. degree in computer science and technology from Hunan University, China, in 2011. He is currently a professor with the College of Cyber Security, Jinan University, China. He was a visiting scholar at New Jersey Institute of Technology, USA, in 2015, and was a visiting professor at Sungkyunkwan University, Korea, in 2016. He serves as a managing editor for IJAACS. His research interests include AI security, cloud computing security, and digital forensic.

Xin Liao (Senior Member, IEEE) received the B.E. and Ph.D. degrees in information security from the Beijing University of Posts and Telecommunications in 2007 and 2012, respectively. He is currently a Professor with the College of Cyber Science and Technology, Hunan University, where he also serves as the Vice Dean. His current research interests include multimedia forensics, AI security, and watermarking. He is the Vice Chair of Technical Committee (TC) on Multimedia Security and Forensics of Asia-Pacific Signal and Information Processing Association. He is serving as an Associate Editor for the IEEE Signal Processing Magazine.

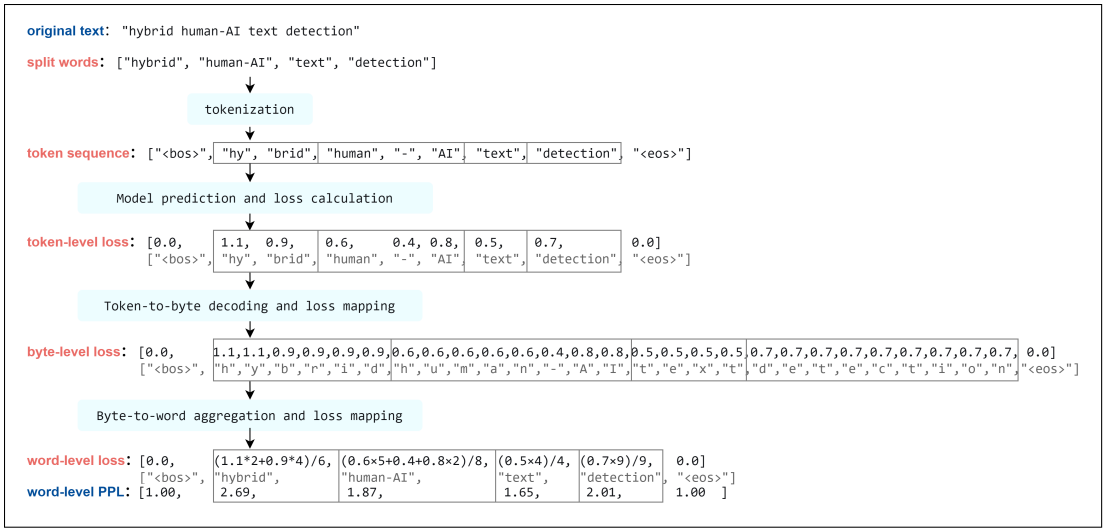


Fig. 7. The workflow for computing word-level perplexity. Each black rectangle denotes a single word; all numerical values are illustrative and do not originate from actual experiments.

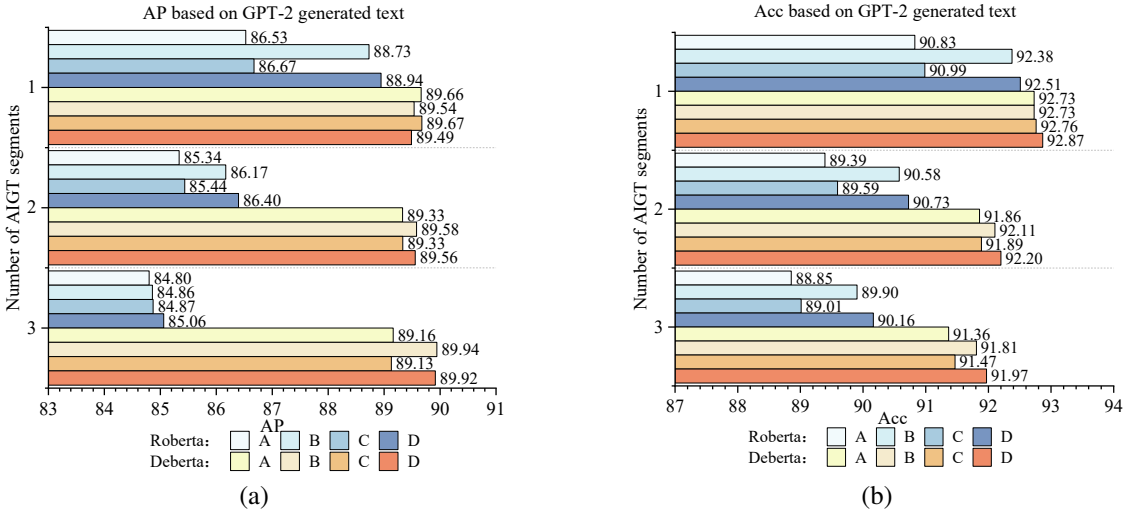


Fig. 8. The results under diverse segmentation conditions on the GoodNews generated by GPT-2: (a)AP and (b)Acc

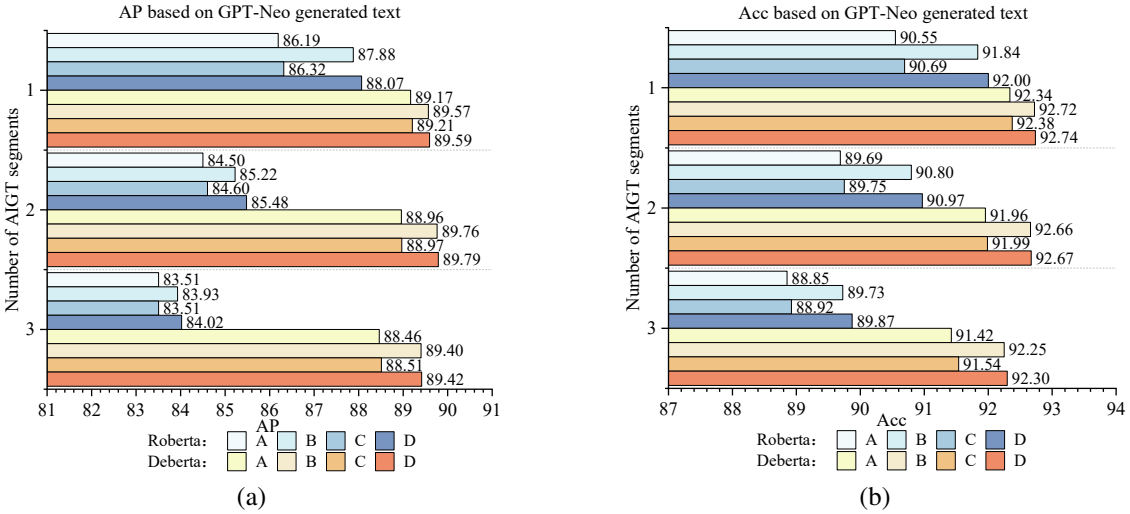


Fig. 9. The results under diverse segmentation conditions on the GoodNews generated by GPT-Neo: (a)AP and (b)Acc

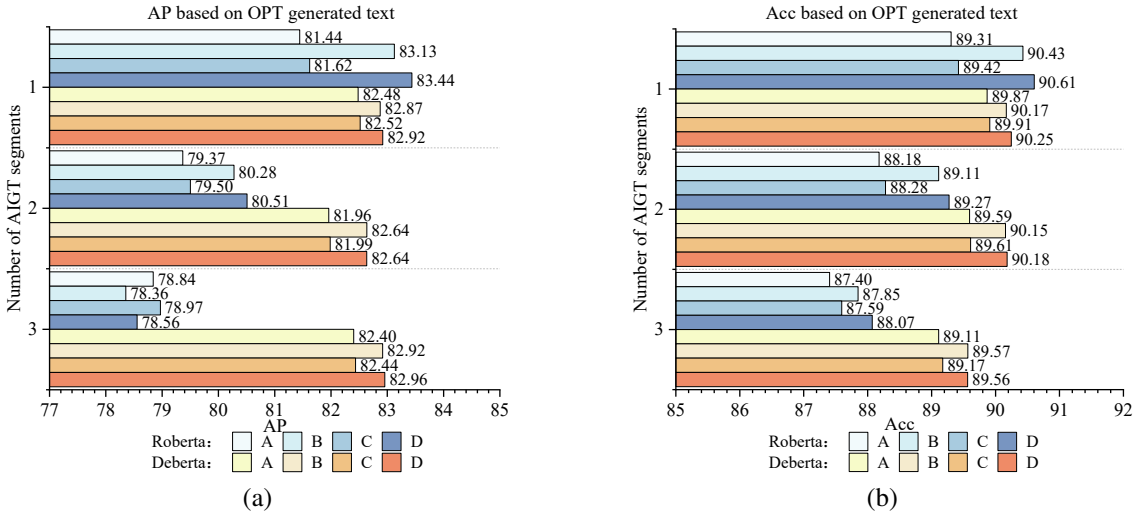


Fig. 10. The results under diverse segmentation conditions on the GoodNews generated by OPT: (a)AP and (b)Acc

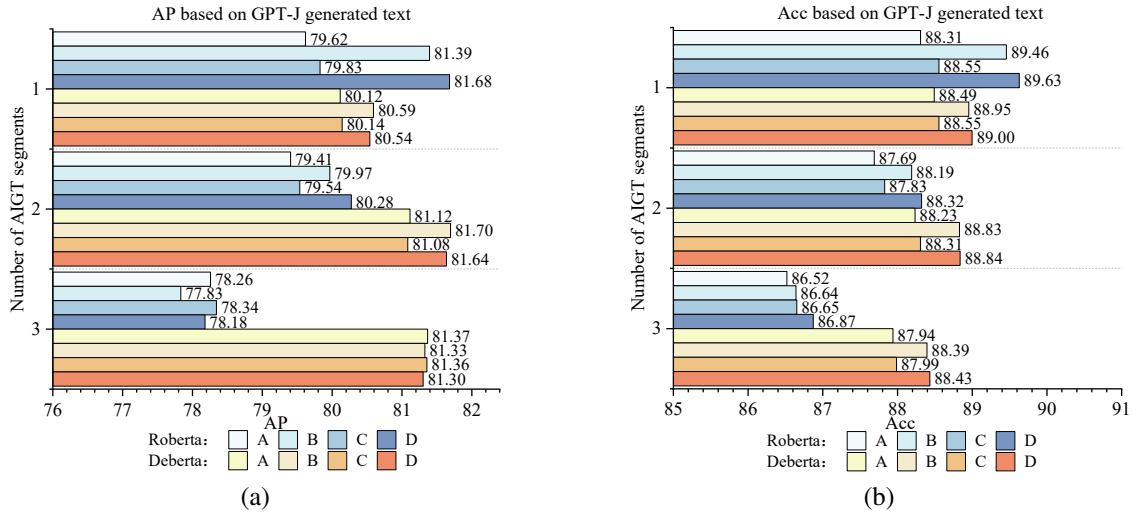


Fig. 11. The results under diverse segmentation conditions on the GoodNews generated by GPT-J: (a)AP and (b)Acc

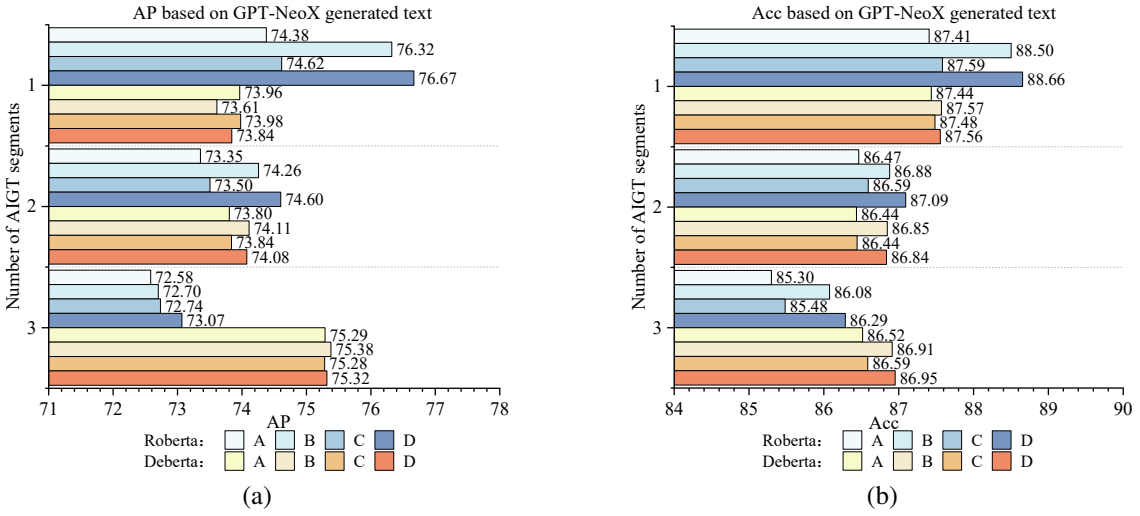


Fig. 12. The results under diverse segmentation conditions on the GoodNews generated by GPT-NeoX: (a)AP and (b)Acc