Are AI-Generated Text Detectors Robust to Adversarial Perturbations?

Guanhua Huang[1*], **Yuchen Zhang**[2†], **Zhe Li**[2], **Yongjian You**[2],
Mingze Wang[3] and **Zhouwang Yang**[1†]
[1]University of Science and Technology of China [2]Bytedance [3]Peking University
guanhuahuang@mail.ustc.edu.cn
{zhangyuchen.zyc, lizhe.2023, youyongjian.cc}@bytedance.com
mingzewang@stu.pku.edu.cn, yangzw@ustc.edu.cn

Abstract

The widespread use of large language models (LLMs) has sparked concerns about the potential misuse of AI-generated text, as these models can produce content that closely resembles human-generated text. Current detectors for AI-generated text (AIGT) lack robustness against adversarial perturbations, with even minor changes in characters or words causing a reversal in distinguishing between human-created and AI-generated text. This paper investigates the robustness of existing AIGT detection methods and introduces a novel detector, the Siamese Calibrated Reconstruction Network (SCRN). The SCRN employs a reconstruction network to add and remove noise from text, extracting a semantic representation that is robust to local perturbations. We also propose a siamese calibration technique to train the model to make equally confident predictions under different noise, which improves the model's robustness against adversarial perturbations. Experiments on four publicly available datasets show that the SCRN outperforms all baseline methods, achieving 6.5%-18.25% absolute accuracy improvement over the best baseline method under adversarial attacks. Moreover, it exhibits superior generalizability in cross-domain, cross-genre, and mixed-source scenarios. The code is available at https://github.com/CarlanLark/Robust-AIGC-Detector.

1 Introduction

Large Language Models (LLMs) such as GPT-4 (Achiam et al., 2023) have shown great promise in producing text that closely mimics human language (Wang et al., 2023b; Chiang and Lee, 2023; Park et al., 2023). However, concerns about the misuse of AI-generated text (AIGT) have arisen in various areas, including the spread of fake news (Hanley and Durumeric, 2023), academic dishonesty (Perkins et al., 2023), and gender bias (Wan et al., 2023). To tackle these issues, various AIGT detection methods have been developed, using statistical features from language models and text features from different model architectures and training approaches (Solaiman et al., 2019; Gehrmann et al., 2019; Mitchell et al., 2023; Guo et al., 2023).

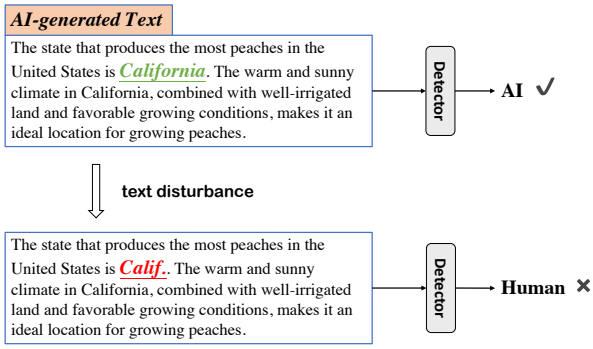

Figure 1: An example of adversarial perturbation to a RoBERTa-based AIGT detector.

Current AI-generated text (AIGT) detectors can effectively identify AI-generated text but struggle with minor adversarial perturbations, such as word substitutions or character swapping (Peng et al., 2023; Cai and Cui, 2023). Small changes that do not change the original text's meaning can cause these detectors to fail. Figure 1 shows a concrete example: a RoBERTa-based AIGT detector can be fooled into classifying AI-generated text as human-written by simply abbreviating "California" to "Calif." This example underscores the limitations of relying solely on token-level features. Therefore, developing robust AIGT detection methods that rely on high-level features is crucial to counteract adversarial perturbation attacks.

To address these challenges, we introduce the Siamese Calibrated Reconstruction Network (SCRN), which consists of an encoder, a reconstruction network, and a classification head. The model first converts input texts into token represen-

[*]Work done during ByteDance Research internship.
[†]Corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6005–6024

tations, then introduces random Gaussian noise to simulate a perturbation attack. The reconstruction network, acting as a denoising auto-encoder (Vincent et al., 2008), aims to remove this noise and recover the original representations. The classification head processes these denoised features to produce the final result. During training, we optimize both classification and reconstruction losses, encouraging the model to learn representations that are resilient to random input perturbations.

Empirically, we observe that a model trained for robustness against random perturbations may not necessarily be robust against adversarial perturbations. To address this issue, we introduce a training technique called siamese calibration. During training, the model generates two classification results using two independent sets of random noise. The training procedure aims to minimize the symmetric Kullback–Leibler (KL) divergence between the two output probability distributions. Since KL divergence is sensitive to changes in probabilities at all confidence levels, the model can incur a significant loss even when it makes consistently correct predictions but with varying confidence levels due to different noise. This stronger constraint forces the model to make equally confident predictions regardless of the noise. In experiments, we find that this approach encourages the model to rely more on high-level contextual features, thereby significantly enhancing its robustness against adversarial attacks.

Our contributions are as follows:

(1) We introduce a reconstruction network that enhances the model's robustness by promoting the learning of resilient representations under token-level perturbations.

(2) We propose a siamese calibration technique that trains the model to make predictions with consistent confidence levels for various random perturbations, which improves its robustness against adversarial attacks.

(3) We establish a comprehensive benchmark for assessing the robustness of AIGT detection methods against a range of adversarial perturbations, including word-level and character-level substitution, deletion, insertion, and swapping. This benchmark encompasses a wide variety of detectors, such as metric-based and model-based detectors, trained using different methods. We evaluate these detectors on four publicly available datasets to test their robustness in in-domain, cross-domain, cross-genre, and mixed-source scenarios.

(4) Our experiments on the benchmark show that SCRN significantly outperforms all baselines in terms of robustness, achieving higher accuracy under adversarial perturbation attacks. Specifically, our method improves over the best baseline method by 11.25, 18.25, 14.5, and 15.75 absolute points of accuracy under attack in in-domain, cross-domain, cross-genre, and mixed-source scenarios, respectively.

2 Related Work

In recent years, Large Language Models (LLMs) like GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) have shown impressive performance in various natural language generation tasks (Kamalloo et al., 2023; Wang et al., 2023b; Cheng et al., 2023b; Chiang and Lee, 2023; Park et al., 2023; Qin et al., 2023). However, the advent of more advanced models such as GPT-4 (Achiam et al., 2023) has raised concerns about the potential misuse of AI-generated texts (AIGT) in areas like fake news (Hanley and Durumeric, 2023; Zhou et al., 2023), academic cheating (Perkins et al., 2023; Foltynek et al., 2023), and ingroup bias (Wan et al., 2023; Gallegos et al., 2023). This highlights the importance of robust detection mechanisms to ensure the security and reliability of applications using LLMs.

To differentiate between human-written and AI-generated texts, various AIGT detection methods have been developed (Gehrmann et al., 2019; Ippolito et al., 2020; Uchendu et al., 2021; Guo et al., 2023; Mitchell et al., 2023). These methods fall into two categories: metric-based and model-based. Metric-based methods use a language model to generate scores for the text and create statistical features from them, such as probability score (Solaiman et al., 2019), rank score (Mitchell et al., 2023), and entropy score (Gehrmann et al., 2019). Model-based methods, on the other hand, employ neural network architectures and supervised learning to train detectors end-to-end using labeled text. For example, OpenAI trained a RoBERTa model to detect GPT-2-generated text (Solaiman et al., 2019), while (Guo et al., 2023) developed a ChatGPT detector based on question-answer text from various domains.

However, research shows that both metric-based and model-based detectors are susceptible to adversarial perturbations (Cai and Cui, 2023; Peng et al., 2023), such as synonym replacement (Ren et al.,

2019; Jin et al., 2020). This means that minor word changes can lead to incorrect classification of AI-generated texts in various fields, including news, education, and finance.(Peng et al., 2023) While robust methods have been introduced in related areas like sentiment analysis (Wang et al., 2023c) and speech recognition (Cheng et al., 2023a), a comprehensive analysis of the resilience of AIGT detectors against adversarial perturbations remains lacking. In this work, we aim to explore the robustness of existing detectors against adversarial perturbations, both in-domain and out-of-domain.

3 Siamese Calibrated Reconstruction Network

3.1 Model Architecture

Encoder Given a dataset $\mathcal{D} = \{(x, y)\}$, where $x = (w_1, w_2, ..., w_n)$ is the input text with n tokens and y denotes the binary label (human or AI), we use a pre-trained RoBERTa (Liu et al., 2019) as the encoder to obtain the text representation:

$$[h_1, h_2, ..., h_n] = \text{RoBERTa}(w_1, w_2, ..., w_n)$$

where $h_i \in \mathbb{R}^d$ are the encoded tokens.

Reconstruction Network To make the detector more robust to text perturbations, we simulate an actual perturbation by adding random noise to each h_i, then utilize a reconstruction network to remove the noise. To inject noise, we split the token representation into a semantic term and a perturbation term, where the former retains the semantic meaning and the latter contains noise. This approach is inspired by (John et al., 2019), which separates the text representation into semantic and style terms to control text style.

Specifically, for the i-th token representation $h_i \in \mathbb{R}^d$, we use an MLP encoder to map it to a lower-dimensional latent space:

$$z_i = \text{MLP}^{(enc)}(h_i)$$

where $z_i \in \mathbb{R}^{d_z}$ is the i-th latent representation. We define a semantic term $z_i^{(s)}$ and a perturbation term $z_i^{(p)}$ based on z_i as follows:

$$z_i^{(s)} = W^{(s)} z_i + b^{(s)}$$
$$z_i^{(p)} = W^{(p)} z_i + b^{(p)}$$

Here, $W^{(s)} \in \mathbb{R}^{d_z \times d_z}, W^{(p)} \in \mathbb{R}^{d_z \times 1}, b^{(s)} \in \mathbb{R}^{d_z}, b^{(p)} \in \mathbb{R}$ are trainable parameters.

Then we combine the two terms to define a noisy latent representation:

$$\tilde{z}_i = z_i^{(s)} + \epsilon \cdot z_i^{(p)} \cdot \mathbb{I}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ is a standard Gaussian noise and $\mathbb{I} \in \mathbb{R}^{d_z}$ is a scalar vector with each entry equal to 1.

To control the numerical scale of latent representations, we introduce a regularization term:

$$\mathcal{L}_{\text{reg}}(x) = \frac{1}{n} \sum_{i=1}^{n} \left(\|z_i^{(s)}\|_2^2 + |z_i^{(p)}|^2 - \alpha \cdot \log(|z_i^{(p)}|) \right)$$

Here, the first two terms serve to penalize the large numerical scale of the latent representations, while the third term aims to prevent the noise scale from decreasing to zero. The hyperparameter α allows for adjusting the degree of penalty applied.

Finally, an MLP decoder is applied to reconstruct the original token representation based on the noisy latents:

$$h_i^{(re)} = \text{MLP}^{(dec)}(\tilde{z}_i)$$

where $h_i^{(re)} \in \mathbb{R}^d$ represents the reconstructed i-th token representation. The reconstruction error is the mean square error between the reconstructed representation and the original one:

$$\mathcal{L}_{\text{mse}}(x) = \frac{1}{n} \sum_{i=1}^{n} ||h_i^{(re)} - h_i||_2^2$$

The final reconstruction loss is the sum of the reconstruction error and the regularization term, where β is a hyperparameter.

$$\mathcal{L}_{\text{re}} = -\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} (\mathcal{L}_{\text{mse}}(x) + \beta \cdot \mathcal{L}_{\text{reg}}(x))$$

Classification Head After obtaining the reconstructed token representations, we use a max pooling layer to extract the feature of the final classification:

$$h^{(cls)} = \text{MaxPooling}([h_1^{(re)}, ..., h_n^{(re)}]).$$

Then we predict the label $\hat{y}$ using a MLP classifier. The classification loss is the standard cross-entropy loss, denoted by $\mathcal{L}_{\text{cls}}$.

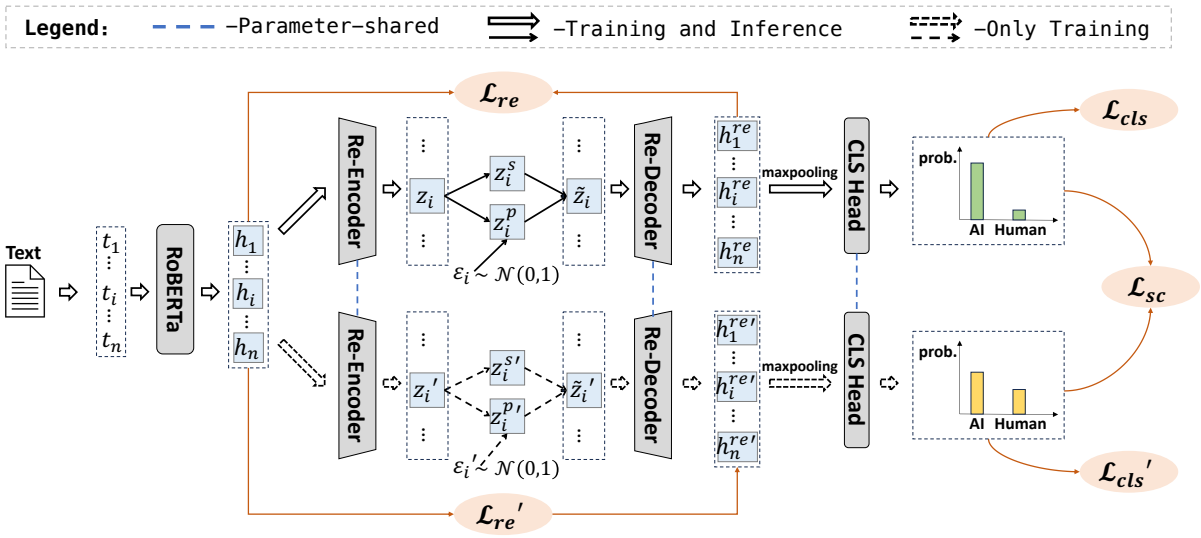

Figure 2: The architecture of SCRN. The input is first encoded by a pre-trained RoBERTa encoder. Then the representation is mapped to a lower-dimensional space by the Re-Encoder to construct the semantic term and the perturbation term, based on which the representation is reconstructed by the Re-Decoder. The denoised representation is used to predict class distributions. Finally, a discrepancy loss is minimized to calibrate the class distributions of two parameter-shared branches.

3.2 Siamese Calibration

The reconstruction loss helps the detector learn robust representations against random perturbations, but it does not guarantee robustness against targeted adversarial perturbations. Empirically, we find that the model, optimized with both reconstruction and classification losses, remains vulnerable to adversarial attacks.

To address this issue, we propose a siamese calibration training strategy. This strategy aims to minimize the symmetric Kullback-Leibler (KL) divergence between the outputs of two inference branches, each subjected to independent random noises, given the same input. Specifically, let $P(x, \epsilon)$ be the predicted class distribution for input x with noise ϵ. The symmetric divergence is then defined as the average of $D_{\mathrm{KL}}(P(x,\epsilon)||P(x,\epsilon'))$ and $D_{\mathrm{KL}}(P(x,\epsilon')||P(x,\epsilon))$, where ϵ and ϵ' are independent copies of random noise, and D_{KL} represents the Kullback–Leibler (KL) divergence. Unlike the cross-entropy loss, which is negligible for correct predictions with high confidence, the symmetric divergence is sensitive to all confidence levels. For instance, consider the correct label is AI. If the first branch predicts AI with a probability of 0.99 and the second branch predicts AI with a probability of $1-\delta$ perturbated by a different random noise, where $\delta < 0.01$, the cross-entropy loss for both branches is less than 0.01, providing little incentive for further optimization. However, the symmetric divergence between these two distributions can approach infinity as $\delta \to 0$. We define the *siamese calibration loss*, denoted as $\mathcal{L}_{\mathrm{sc}}$, as the symmetric divergence across all training inputs $x \in D$. This loss specifically penalizes inconsistent confidence levels, imposing a stricter requirement than the cross-entropy loss. By minimizing this loss, the detector is encouraged to focus on high-level contextual features that are less susceptible to token perturbations.

For training the SCRN model, we make a weighted summing over the above three losses:

$$\mathcal{L}_{\mathrm{all}} = \lambda_1(\mathcal{L}_{\mathrm{cls}} + \mathcal{L}'_{\mathrm{cls}}) + \lambda_2(\mathcal{L}_{\mathrm{re}} + \mathcal{L}'_{\mathrm{re}}) + \lambda_3\mathcal{L}_{\mathrm{sc}}$$

where $\mathcal{L}_{\mathrm{cls}}, \mathcal{L}'_{\mathrm{cls}}$ are classification losses of two branches, $\mathcal{L}_{\mathrm{re}}, \mathcal{L}'_{\mathrm{re}}$ are reconstruction losses, and $\mathcal{L}_{\mathrm{sc}}$ is siamese calibration loss. $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters.

Siamese calibration is only applicable during training. During inference, the detector generates a single copy of random noise and produces a single prediction. Training with siamese calibration significantly enhances the model's prediction consistency during inference. Analysis in Appendix A.7 shows that the fluctuation in inference robustness between two independent branches becomes negligible after implementing siamese calibration.

Scenarios	Train Set	Test Set	Num. Train	Num. Test	Num. Attack
In-domain	HC3 train	HC3 test	76,905	8,544	400
Cross-domain	HC3 train	TruthfulQA	76,905	1,634	400
Cross-genre	HC3 train	GhostBuster	76,905	6,000	400
Mixed-source	SeqXGPT-Bench train	SeqXGPT-Bench test	10,800	1,200	400

Table 1: Statistics of datasets for different AIGT detection scenarios.

4 Experiments

4.1 Experimental Setup

Datasets To assess the robustness of AIGT detectors to adversarial perturbations, we conduct experiments on four public datasets: **HC3**(Guo et al., 2023), **TruthfulQA**(He et al., 2023), **Ghostbuster**(Verma et al., 2023), and **SeqXGPT-Bench**(Wang et al., 2023a). These datasets encompass a wide range of AIGT scenarios, including in-domain, cross-domain, cross-genre, and mixed-source. The datasets and number of samples used for training and evaluating AIGT detectors are listed in Table 1. For dataset details, see Appendix A.1.

Metrics Following the approach of (Wang et al., 2023c), we assess model accuracy and robustness using four metrics: (1) Original Accuracy (**OA**% ↑) measures the model's raw accuracy without adversarial perturbations. (2) Accuracy Under Attack (**AUA**% ↑) quantifies the model's accuracy on adversarially perturbed text. (3) Attack Success Rate (**ASR**% ↓) indicates the percentage of test samples successfully fooled by the attacker. (4) Average Number of Queries (**ANQ** ↑) represents the average number of adversarial attack queries on test samples, with higher values indicating a more robust model. The symbols ↑ and ↓ denote that higher and lower values are better, respectively. Additionally, we report traditional **Precision**, **Recall**, and **micro-F1** scores when no attack is performed.

Adversarial Perturbations In our experiments, adversarial perturbations are conducted by three attack methods: **PWWS** (Ren et al., 2019), **DeepWordBug** (Gao et al., 2018), and **DeepWordBug** (Gao et al., 2018). These methods involve various types of adversarial perturbations including character-level and word-level substitution, deletion, and insertion. More details can be found in Appendix A.2

Baselines To create a comprehensive benchmark, we do our best to select a wide range of detector methods. These include metric-based detectors such as **Log-Likelihood**(Solaiman et al., 2019), **Log-Rank**(Mitchell et al., 2023), **Entropy**(Gehrmann et al., 2019), **GLTR**(Gehrmann et al., 2019), and **SeqXGPT**(Wang et al., 2023a), along with model-based detectors like **Bert**(Devlin et al., 2019), **Roberta**(Liu et al., 2019), **Deberta**(He et al., 2020), and **ChatGPT-Detector**(Guo et al., 2023). We also implemented recent methods for enhancing classification robustness in other NLP tasks, such as **Flooding**(Ishida et al., 2020), **RDrop**(Wu et al., 2021), **RanMASK**(Zeng et al., 2023), and **RMLM** (Wang et al., 2023c). Details can be found in Appendix A.3.

Experiment Settings To ensure a fair comparison of the AIGT detectors, all models were trained on 8 * 32GB NVIDIA V100 GPUs and evaluated on a single 32GB NVIDIA V100 GPU, using the same environment. We utilized the base versions of pre-trained Bert, RoBERTa, and DeBERTa models as employed in the compared detectors. The implemented methods were based on their officially released code, and for the ChatGPT-Detector, we utilized the provided model weight. The hyperparameters of SCRN and more information can be found in Appendix A.4.

4.2 In-domain Robustness

In the in-domain scenario, where the test data comes from the same domain as the training data (HC3), our SCRN model demonstrates superior robustness against various types of adversarial perturbations from three attack methods. As illustrated in Table 3, SCRN achieves a notable improvement in Accuracy Under Attack (AUA), with a 6.5-11.25 absolute increase compared to the best baseline detector. The AUA values for SCRN remain high, exceeding 91.25 under all three attack methods, only marginally lower than its original accuracy without any attacks. Table 3 also reveals that *evasion* attacks (tricking the model into classifying AI-generated answers as human, de-

Methods	P.(AI)	R.(AI)	P.(H.)	R.(H.)	F.(Overall)
Log-Likelihood	95.43	95.61	97.98	97.90	97.18
Log-Rank	96.30	95.87	98.11	98.31	97.54
Entropy	89.90	87.69	94.41	95.47	93.00
GLTR	96.84	95.76	98.06	98.57	97.68
SeqXGPT	97.92	**100.0**	**100.0**	99.03	99.33
Bert	98.28	**100.0**	**100.0**	99.20	99.45
Roberta	99.45	**100.0**	**100.0**	99.74	99.80
Deberta	99.74	**100.0**	**100.0**	99.88	99.92
ChatGPT-Detector	99.03	99.15	99.61	99.56	99.58
Flooding	99.67	**100.0**	**100.0**	99.85	99.89
RDrop	99.67	99.96	99.98	99.85	99.88
RanMASK	87.73	**100.0**	**100.0**	93.58	95.67
RMLM	96.15	**100.0**	**100.0**	96.00	98.00
SCRN	**99.78**	**100.0**	**100.0**	**99.90**	**99.93**

Table 2: Results of **in-domain** AIGT detection **without attack**. P.(AI) and R.(AI) denote Precision and Recall for AI-generated text as positive samples, while P.(H.) and R.(H.) indicate Precision and Recall for human-created text as positive samples.

noted by AI→Human) are easier than *obfuscation* attacks (tricking the model into classifying human-generated answers as AI, denoted by Human→AI) on HC3. This is likely because human-generated answers are diverse (Ma et al., 2023), making it challenging to perturb them all to resemble AI-generated responses.

In the absence of adversarial perturbation, all detectors except RanMASK demonstrate high performance, as illustrated in Table 2. RanMASK's lower accuracy can be attributed to its masking of 30% of the text during both training and inference, resulting in significant information loss.

4.3 Cross-domain Robustness

In the cross-domain setting, where the test data (TruthfulQA) and training data (HC3) come from different domains but share the same genre (question answering), SCRN significantly outperforms other baseline models. As Table 4 and Table 6 demonstrates, SCRN achieves a notable improvement, with at least a +18.25 increase in absolute Accuracy Under Attack (AUA). Although absolute accuracy tends to be lower in cross-domain scenarios, SCRN's margin of lead is even more pronounced.

4.4 Cross-genre Robustness

In the cross-genre setting, the test data (Ghostbuster) is from a different genre than the training data (HC3). Specifically, Ghostbuster comprises news articles, essays, and creative writings, while HC3 is a question-answering dataset. As shown in Table 5 and Table 7, SCRN outperforms all baseline models in overall Accuracy Under Attack (AUA) and Attack Success Rate (ASR).

We observe that SCRN's AUA is lower than some baselines under obfuscation attacks (Human→AI). However, SCRN is substantially more robust under evasion attacks (AI→Human), achieving 71% AUA, while baseline models' scores are near zero. This indicates that in the cross-genre setting, minor perturbations on AI-generated content can easily lead to a "Human" prediction by baseline models, as they are not trained on AI-generated content like news articles, essays, and creative writings. As a by-product, obfuscation attack against these models become very hard. Practically speaking, defending against evasion attacks is more important. SCRN demonstrates balanced robustness, underscoring its generalizability.

It is also noteworthy that RMLM, a model obtained through adversarial training, shows good robustness in the in-domain setting but fails in cross-domain and cross-genre settings. This suggests that merely augmenting the training set with adversarial data is insufficient to enhance the model's robustness against out-of-distribution samples (Wang et al., 2022).

4.5 Mixed-source Robustness

Tables 8 and 9 present the results on the SeqXGPT-Bench dataset, which comprises AI-generated content from various LLMs in mixed-source scenarios. Notably, SCRN shows excellent performance both with and without attacks.

Specifically, model-based detectors like SCRN significantly outperform metric-based detectors in the absence of an attack. However, the accuracy of all models decreases with the PWWS attack. Compared to the best-performing baseline, RMLM, our SCRN achieves a notable improvement of 15.75 in Accuracy Under Attack (AUA), consistent with other settings.

4.6 Ablation Study

In this section, we perform an ablation study to evaluate the effectiveness of our key design choices. To assess the impact of siamese calibration, we train the SCRN model without it, denoted as SCRN-SC. Table 10 shows that the detector's robustness significantly worsens across all four AIGT scenarios. Notably, the accuracy under attack (AUA) of SCRN-SC is lower than the baseline RoBERTa detector (SCRN-SC-R) in the in-domain setting. This decline may be attributed to SCRN-SC's inclusion

	Methods	AI → Human				Human → AI				Overall			
		OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑
PWWS	Log-Likelihood	96.00	0.00	100.00	957.42	**100.00**	99.00	1.00	1223.54	98.00	49.50	49.49	1090.48
	Log-Rank	96.50	0.00	100.00	974.20	99.00	98.50	0.51	1233.61	97.75	49.25	49.62	1103.90
	Entropy	86.00	0.00	100.00	962.97	95.00	79.00	16.84	1112.75	90.50	39.50	56.35	1037.86
	GLTR	95.00	0.00	100.00	986.10	99.00	83.50	15.66	1136.52	97.00	41.75	56.96	1061.31
	SeqXGPT	99.50	11.00	88.94	1368.07	**100.00**	**100.00**	**0.00**	1224.88	99.75	55.50	44.36	1296.47
	BERT	**100.00**	1.50	98.50	1070.74	99.50	98.50	1.01	1211.18	99.75	50.00	49.87	1140.96
	RoBERTa	**100.00**	38.50	61.50	1332.60	**100.00**	99.50	0.50	1223.76	**100.00**	69.00	31.00	1278.18
	DeBERTa	**100.00**	3.00	97.00	1170.89	**100.00**	99.50	0.50	1223.53	**100.00**	51.25	48.75	1197.21
	ChatGPT-Detector	98.00	0.00	100.00	1074.98	**100.00**	99.50	0.50	1224.60	99.00	49.75	49.75	1149.79
	Flooding	**100.00**	23.00	77.00	1422.60	**100.00**	**100.00**	**0.00**	1225.00	**100.00**	61.50	38.50	1323.81
	RDrop	99.50	67.50	32.16	1585.15	**100.00**	**100.00**	**0.00**	**1225.12**	99.75	83.75	16.04	1405.08
	RanMASK	**100.00**	50.00	50.00	1562.84	**100.00**	**100.00**	**0.00**	1245.97	**100.00**	75.00	25.00	1404.40
	RMLM	**100.00**	73.50	26.50	1561.35	**100.00**	98.50	1.50	1216.52	**100.00**	86.00	14.00	1388.94
	SCRN	**100.00**	**94.50**	**5.50**	**1665.53**	**100.00**	**100.00**	**0.00**	1225.02	**100.00**	**97.25**	**2.75**	**1445.28**
Deep-Word-Bug	Log-Likelihood	96.00	0.00	100.00	109.20	**100.00**	**100.00**	**0.00**	306.24	98.00	50.00	48.98	207.72
	Log-Rank	96.50	0.00	100.00	110.93	99.00	99.00	**0.00**	308.03	97.75	49.50	49.36	209.48
	Entropy	86.00	0.00	100.00	108.82	95.00	92.50	2.63	295.70	90.50	46.25	48.90	202.26
	GLTR	95.00	0.00	100.00	114.79	99.00	98.50	0.51	306.09	97.00	49.25	49.23	210.44
	SeqXGPT	99.50	8.00	91.96	139.96	**100.00**	**100.00**	**0.00**	306.16	99.75	54.00	45.86	223.06
	BERT	**100.00**	12.50	87.50	152.30	99.50	98.50	1.01	282.92	99.75	55.50	44.36	217.61
	RoBERTa	**100.00**	53.00	47.00	293.59	**100.00**	**100.00**	**0.00**	302.58	**100.00**	76.50	23.50	298.08
	DeBERTa	**100.00**	32.00	68.00	171.36	**100.00**	**100.00**	**0.00**	295.56	**100.00**	66.00	34.00	233.46
	ChatGPT-Detector	98.00	13.50	86.22	161.07	**100.00**	**100.00**	**0.00**	301.62	99.00	56.75	42.68	231.34
	Flooding	**100.00**	35.00	65.00	175.48	**100.00**	**100.00**	**0.00**	275.71	**100.00**	67.50	32.50	225.59
	RDrop	99.50	60.50	39.20	367.74	**100.00**	**100.00**	**0.00**	306.09	99.75	80.25	19.55	336.92
	RanMASK	**100.00**	59.00	41.00	332.98	**100.00**	**100.00**	**0.00**	**315.78**	**100.00**	79.50	20.50	324.38
	RMLM	**100.00**	66.00	34.00	377.24	**100.00**	**100.00**	**0.00**	308.91	**100.00**	83.00	17.00	343.08
	SCRN	**100.00**	**87.50**	**12.50**	**437.50**	**100.00**	**100.00**	**0.00**	305.93	**100.00**	**93.75**	**6.25**	**371.72**
Pruthi	Log-Likelihood	96.00	1.00	98.96	9000.48	**100.00**	**100.00**	**0.00**	9519.34	98.00	50.50	48.47	9259.91
	Log-Rank	96.50	1.00	98.96	10084.75	99.00	99.00	**0.00**	9557.18	97.75	50.00	48.85	9820.96
	Entropy	86.00	0.00	100.00	7920.77	95.00	90.50	4.74	9085.92	90.50	45.25	50.00	8503.35
	GLTR	95.00	1.00	98.95	10321.49	99.00	95.00	4.04	9507.22	97.00	48.00	50.52	9914.36
	SeqXGPT	99.50	1.00	98.99	10505.46	**100.00**	**100.00**	**0.00**	9609.78	99.75	50.50	49.37	10057.62
	BERT	**100.00**	21.00	79.00	17460.63	99.50	98.00	1.51	9502.22	99.75	59.50	40.35	13481.42
	RoBERTa	**100.00**	57.00	43.00	20431.19	**100.00**	**100.00**	**0.00**	9561.62	**100.00**	78.50	21.50	14996.40
	DeBERTa	**100.00**	34.50	65.50	17338.08	**100.00**	99.50	0.50	9606.24	**100.00**	67.00	33.00	13472.16
	ChatGPT-Detector	98.00	36.50	62.76	18563.57	**100.00**	99.50	0.50	9591.88	99.00	68.00	31.31	14077.72
	Flooding	**100.00**	68.50	31.50	20823.59	**100.00**	**100.00**	**0.00**	9540.92	**100.00**	84.25	15.75	15182.26
	RDrop	99.50	69.00	30.65	20132.45	**100.00**	99.50	0.50	9516.64	99.75	84.25	15.54	14824.54
	RanMASK	**100.00**	68.50	31.50	21052.49	**100.00**	98.00	2.00	**9748.31**	**100.00**	83.25	16.75	**15400.40**
	RMLM	**100.00**	71.50	28.50	20949.12	**100.00**	98.00	2.00	9373.92	**100.00**	84.75	15.25	15161.52
	SCRN	**100.00**	**82.50**	**17.50**	**21122.83**	**100.00**	**100.00**	**0.00**	9540.20	**100.00**	**91.25**	**8.75**	15331.52

Table 3: Results of **in-domain** AIGT detection **under different adversarial attack methods**.

	P.(AI)	R.(AI)	P.(H.)	R.(H.)	F.(Overall)
Log-Likelihood	95.88	76.99	80.78	96.67	86.71
Log-Rank	94.86	72.21	77.57	96.08	83.92
Entropy	97.56	48.84	65.88	98.78	72.06
GLTR	87.61	71.85	76.14	89.84	80.69
SeqXGPT	95.03	88.86	89.54	95.35	92.10
Bert	**99.54**	80.05	83.32	**99.63**	89.74
Roberta	82.77	87.03	86.32	86.32	84.45
Deberta	97.99	89.35	90.21	98.16	93.75
ChatGPT-Detector	97.94	**92.90**	**93.25**	98.04	**95.47**
Flooding	84.19	88.00	87.44	83.48	85.73
RDrop	90.43	87.88	88.21	90.70	89.29
RanMASK	86.54	85.80	85.92	86.66	86.23
RMLM	97.61	80.05	83.09	98.04	88.96
SCRN	98.71	83.97	86.05	98.90	91.38

Table 4: Results of **cross-domain** AIGT detection on SeqXGPT-Bench dataset **without attack**.

	P.(AI)	R.(AI)	P.(H.)	R.(H.)	F.(Overall)
Log-Likelihood	92.60	75.03	79.01	94.00	84.38
Log-Rank	95.58	63.43	72.64	97.07	79.68
Entropy	72.45	75.03	74.11	71.47	73.24
GLTR	**96.36**	53.77	67.94	**97.97**	74.63
SeqXGPT	86.32	84.37	84.71	86.63	**85.50**
Bert	89.82	59.10	69.52	93.30	75.48
Roberta	83.07	83.43	83.36	83.00	83.22
Deberta	79.51	90.67	89.14	76.63	83.57
ChatGPT-Detector	86.46	69.40	69.08	90.70	74.42
Flooding	83.75	87.60	87.00	83.00	85.29
RDrop	77.05	94.43	**92.81**	71.87	82.93
RanMASK	89.84	68.10	74.32	92.30	79.90
RMLM	89.90	57.83	68.92	93.50	74.87
SCRN	73.61	**94.67**	92.53	66.07	79.96

Table 5: Results of **cross-genre** AIGT detection on SeqXGPT-Bench dataset **without attack**.

of random noise during inference, which, without siamese calibration, leads to unstable predictions

	Methods	AI → Human				Human → AI				Overall			
		OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑
PWWS	Log-Likelihood	67.00	0.00	100.00	381.32	97.00	97.00	0.00	89.96	82.00	48.50	40.85	235.64
	Log-Rank	72.00	0.00	100.00	374.13	95.50	95.00	0.52	89.63	83.75	47.50	43.28	231.88
	Entropy	43.50	0.00	100.00	433.91	98.00	80.00	18.37	83.96	70.75	40.00	43.46	258.94
	GLTR	66.50	0.00	100.00	376.47	88.50	56.50	36.16	79.80	77.50	28.25	63.55	228.14
	SeqXGPT	93.00	4.00	95.70	397.55	98.50	94.00	4.56	97.64	95.75	49.00	48.83	247.60
	BERT	80.00	0.00	100.00	384.23	99.50	99.50	0.00	89.24	89.75	49.75	44.57	236.74
	RoBERTa	90.50	6.50	92.82	410.38	75.50	74.00	1.99	**99.81**	83.00	40.25	51.51	255.10
	DeBERTa	91.50	1.00	98.91	381.80	98.00	97.50	0.51	88.78	94.75	49.25	48.02	235.29
	ChatGPT-Detector	**96.00**	1.00	98.96	364.00	98.50	88.50	10.15	85.93	**97.25**	44.75	53.98	224.96
	Flooding	90.00	0.00	100.00	387.93	78.00	74.50	4.49	101.59	84.00	37.25	55.65	244.76
	RDrop	89.50	6.00	93.30	429.06	89.00	73.50	17.42	91.10	89.25	39.75	55.46	260.08
	RanMASK	89.00	1.00	98.88	406.12	81.00	78.00	3.70	98.26	85.00	39.50	53.53	252.19
	RMLM	81.50	5.50	93.25	426.98	98.00	98.00	0.00	91.56	89.75	51.75	42.34	259.27
	SCRN	86.50	**40.50**	**53.18**	**551.20**	**99.50**	**99.50**	**0.00**	89.24	93.00	**70.00**	**24.73**	**320.22**

Table 6: Results of **cross-domain** AIGT detection **under PWWS attack**. More results refer to Appendix A.9.

	Methods	AI → Human				Human → AI				Overall			
		OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑
PWWS	Log-Likelihood	62.00	0.00	100.00	2700.46	**97.50**	**96.50**	**1.03**	**6077.26**	79.75	48.25	39.50	4388.86
	Log-Rank	64.50	0.00	100.00	2734.98	97.50	95.50	2.05	6054.48	81.00	47.75	41.05	4394.73
	Entropy	77.50	0.00	100.00	2783.14	74.00	34.00	54.05	5352.47	75.75	17.00	77.56	4067.80
	GLTR	50.50	0.00	100.00	2696.80	97.50	67.50	30.77	5476.04	74.00	33.75	54.39	4086.42
	SeqXGPT	85.50	0.00	100.00	2712.97	88.00	65.50	25.57	5776.35	**86.75**	32.75	62.25	4244.66
	BERT	57.00	0.00	100.00	2692.61	95.50	75.00	21.47	5619.45	76.25	37.50	50.82	4156.03
	RoBERTa	82.00	0.00	100.00	2655.43	83.00	59.00	28.92	5522.05	82.50	29.50	64.24	4088.74
	DeBERTa	90.00	0.00	100.00	2763.66	77.50	53.50	30.97	5329.64	83.75	26.75	68.06	4046.65
	ChatGPT-Detector	58.50	0.00	100.00	2606.75	93.00	73.00	21.51	5827.88	75.75	36.50	51.82	4217.32
	Flooding	87.50	0.00	100.00	2733.18	82.50	58.00	29.70	5447.84	85.00	29.00	65.88	4090.51
	RDrop	95.00	10.00	89.47	3155.59	73.00	65.00	10.96	5973.84	84.00	37.50	55.36	4564.72
	RanMASK	67.00	2.00	97.01	2667.19	87.00	75.00	13.79	5433.82	77.00	38.50	50.00	4050.50
	RMLM	58.50	9.50	83.76	3397.99	92.00	72.50	21.20	5440.61	75.25	41.00	45.51	4419.30
	SCRN	**94.50**	**71.00**	**24.87**	**4419.16**	70.50	54.50	22.70	5725.79	82.50	**62.75**	**23.94**	**5072.48**

Table 7: Results of **cross-genre** AIGT detection **under PWWS attack**. More results refer to Appendix A.10.

	Methods	AI → Human				Human → AI				Overall			
		OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑
PWWS	Log-Likelihood	72.00	0.50	99.31	1281.86	62.00	53.50	13.71	1667.91	67.00	27.00	59.70	1474.88
	Log-Rank	73.50	0.50	99.32	1286.24	62.50	56.00	10.40	1697.20	68.00	28.25	58.46	1491.72
	Entropy	63.00	0.00	100.00	1239.29	55.50	27.50	50.45	1396.39	59.25	13.75	76.79	1317.84
	GLTR	76.50	0.00	100.00	1260.99	67.50	19.00	71.85	1285.64	72.00	9.50	86.81	1273.32
	SeqXGPT	**96.50**	65.00	32.64	1867.81	96.00	70.00	27.08	1893.98	**96.25**	67.50	29.87	1880.90
	BERT	90.50	1.00	98.90	1204.52	90.00	59.00	34.44	1815.44	90.25	30.00	66.76	1509.98
	RoBERTa	95.50	64.50	32.46	1840.19	93.00	62.50	32.80	1729.72	94.25	63.50	32.63	1784.96
	DeBERTa	95.50	54.50	42.93	1764.94	96.00	80.00	16.67	1940.47	95.75	67.25	29.77	1852.70
	Flooding	96.00	60.50	36.98	1800.01	95.50	53.00	44.50	1610.45	95.75	56.75	40.73	1705.23
	RDrop	**96.50**	69.00	28.50	1819.95	95.00	70.00	26.32	1815.62	95.75	69.50	27.42	1817.78
	RanMASK	94.00	60.00	36.17	1784.11	86.00	71.00	17.44	1715.72	90.00	65.50	27.22	1749.92
	RMLM	91.00	69.00	24.18	1879.96	91.50	78.00	14.75	1986.50	91.25	73.50	19.45	1933.23
	SCRN	95.00	**87.00**	**8.42**	**1986.98**	**96.00**	**91.50**	**4.69**	**2099.91**	95.50	**89.25**	**6.54**	**2043.44**

Table 8: Results of **mixed-source** AIGT detection **under PWWS attack**. The AI-generated texts are from five sources: GPT-2, GPT-Neo, GPT-J, LLaMa, and GPT-3. More results refer to Appendix A.11.

under adversarial attacks. These findings underscore the importance of siamese calibration.

To examine the role of the reconstruction network, we replace it with a simple dropout layer, resulting in the SCRN-R model. As depicted in Table 10, SCRN-R experiences a decrease in Accuracy Under Attack ranging from 18.0 to 33.25 across the four scenarios. This decline occurs because the dropout layer merely omits text information without simulating adversarial perturbations, which involve more complex editing actions such as substitution, insertion, and deletion.

Regarding the noise of the reconstruction network, completely removing the noise ϵ (SCRN-ϵ) leads to a significant decrease in the detector's robustness, particularly in cross-domain scenarios of

	P.(AI)	R.(AI)	P.(H.)	R.(H.)	F.(Overall)
Log-Likelihood	64.89	69.00	66.90	62.67	65.80
Log-Rank	65.69	70.83	68.35	63.00	66.87
Entropy	57.26	59.17	57.76	55.83	57.49
GLTR	69.01	73.50	71.66	67.00	70.22
SeqXGPT	95.03	**97.39**	**97.32**	94.93	**96.15**
Bert	86.76	90.67	90.23	86.17	88.50
Roberta	94.77	96.66	96.60	94.67	95.67
Deberta	**95.99**	95.83	95.84	**96.00**	95.92
Flooding	94.91	96.33	96.28	94.83	95.58
RDrop	94.63	97.00	96.92	94.50	95.75
RanMASK	78.11	95.17	93.82	73.33	84.06
RMLM	85.34	92.17	91.49	84.17	88.15
SCRN	95.69	96.17	96.15	95.67	95.92

Table 9: Results of **mixed-source** AIGT detection on SeqXGPT-Bench dataset **without attack**. The AI-generated texts are from five sources: GPT-2, GPT-Neo, GPT-J, LLaMa, and GPT-3.

29.5 AUA drop. Furthermore, when we dropped the regularization loss $\mathcal{L}_{reg}$, although the noise still existed, it degenerated during the optimization of $\mathcal{L}_{mse}$. The AUA scores of SCRN-$\mathcal{L}_{reg}$ demonstrated that maintaining a degenerated noise showed better robustness compared to completely removing the noise (SCRN-ϵ). However, SCRN-$\mathcal{L}_{reg}$ performed significantly worse than SCRN, which highlights the necessity of the regularization loss $\mathcal{L}_{reg}$ in preserving the noise's effectiveness and maintaining the desired robustness.

		OA ↑	AUA ↑	ASR ↓	ANQ ↑
In-domain	SCRN	100.0	97.25	2.75	1445.28
	-SC	99.25	50.00	49.62	1067.74
	-R	99.50	79.25	20.35	1373.16
	-ϵ	100.00	92.00	8.00	1432.34
	-$\mathcal{L}_{reg}$	100.00	96.75	3.25	1444.54
	-SC-R	100.0	69.00	31.00	1278.18
Cross-domain	SCRN	93.00	70.00	24.73	320.22
	-SC	87.00	42.50	51.15	228.30
	-R	88.00	36.75	58.24	222.51
	-ϵ	82.50	40.50	50.91	280.56
	-$\mathcal{L}_{reg}$	87.25	57.50	34.10	311.72
	-SC-R	83.00	40.25	51.51	255.10
Cross-genre	SCRN	82.50	62.75	23.94	5072.48
	-SC	86.25	41.75	51.59	4318.61
	-R	84.50	35.00	58.58	4359.77
	-ϵ	83.00	43.50	47.59	4782.50
	-$\mathcal{L}_{reg}$	79.50	59.50	25.16	4920.92
	-SC-R	82.50	29.50	64.24	4088.74
Mixed-source	SCRN	95.50	89.25	6.54	2043.44
	-SC	95.00	87.50	7.89	2025.96
	-R	96.00	68.00	29.17	1797.31
	-ϵ	95.75	77.25	19.32	1932.08
	-$\mathcal{L}_{reg}$	96.00	82.75	13.80	1970.64
	-SC-R	94.25	63.50	32.63	1784.96

Table 10: The **ablation study** on four AIGT scenarios **under PWWS attack**.

4.7 More Analysis

We also provide a threshold analysis in Appendix A.5, a comparison of inference speeds in Appendix A.6, an analysis of inference fluctuations in Appendix A.7, and case studies in Appendix A.8.

5 Conclusion

While AI-generated text (AIGT) detection is promising for various applications, it struggles with the robustness of current methods against adversarial perturbations. To tackle this, we introduce the Siamese Calibrated Reconstruction Network (SCRN). SCRN uses a reconstruction network to model text perturbations and employs siamese calibrated training to improve inference robustness. Experiments demonstrate SCRN's effectiveness in defending against adversarial perturbations in four AIGT scenarios, highlighting its practical utility in addressing real-world AIGT detection challenges.

Limitations

Although SCRN demonstrates excellent robust performance across all four scenarios, including in-domain, cross-domain, cross-genre, and mixed-source AIGT detections, it still has several limitations:

(1) We did not consider the text paraphrasing attack (Tulchinskii et al., 2024; Macko et al., 2024) as a form of text perturbation. Our focus was primarily on adversarial perturbations with minor modifications, and we regarded paraphrased text as completely modified. Strictly, if a paraphrased AI-generated text becomes the same as an existing human-created text, it should be assigned a human-created label by the AIGT detector. Future work may conduct a further more detailed analysis of the paraphrased text.

(2) Our experiments mainly focused on English corpora, and while our proposed method is general, we did not explore its performance on multilingual corpora. We leave the detailed analysis of multilingual datasets in future work.

Acknowledgements

The work is supported by the National Key R&D Program of China (Nos. 2022YFA1005201, 2022YFA1005202, 2022YFA1005203) and the NSFC Major Research Plan - Interpretable and General Purpose Next-generation Artificial Intelligence (No. 92270205).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Shuyang Cai and Wanyun Cui. 2023. Evade chatgpt detectors via a single space. *arXiv preprint arXiv:2307.02599*.

Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023a. Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6492–6505.

Xuxin Cheng, Zhihong Zhu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023b. Das-cl: Towards multimodal machine translation via dual-level asymmetric contrastive learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 337–347.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

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

Tomas Foltynek, Sonja Bjelobaba, Irene Glendinning, Zeenath Reza Khan, Rita Santos, Pegi Pavletic, and Július Kravjar. 2023. Enai recommendations on the ethical use of artificial intelligence in education. *International Journal for Educational Integrity*, 19(1):12.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Hans WA Hanley and Zakir Durumeric. 2023. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. *arXiv preprint arXiv:2305.09820*.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. Mgtbench: Benchmarking machine-generated text detection. *arXiv preprint arXiv:2303.14822*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Takashi Ishida, Ikko Yamane, Tomoya Sakai, Gang Niu, and Masashi Sugiyama. 2020. Do we need zero training loss after achieving zero training error? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4604–4614. PMLR.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36.

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

Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. 2023. Ai vs. human–differentiation analysis of scientific content generation. *arXiv*, 2301.

Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason Samuel Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2024. Authorship obfuscation in multilingual machine-generated text detection. *arXiv preprint arXiv:2401.07867*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Xinlin Peng, Ying Zhou, Ben He, Le Sun, and Yingfei Sun. 2023. Hidding the ghostwriters: An adversarial evaluation of AI-generated student essay detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10406–10419, Singapore. Association for Computational Linguistics.

M Perkins, J Roe, D Postma, J McGaughran, D Hickerson, and J Cook. 2023. Game of tones: Faculty detection of gpt-4 generated content in university assessments. arxiv.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Rafael Rivera Soto, Kailin Koch, Aleem Khan, Barry Chen, Marcus Bishop, and Nicholas Andrews. 2024. Few-shot detection of machine-generated text using style representations. *arXiv preprint arXiv:2401.06712*.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2024. Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, 36.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2023. Ghostbuster: Detecting text ghostwritten by large language models. *arXiv preprint arXiv:2305.15047*.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023a. SeqXGPT: Sentence-level AI-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1144–1156, Singapore. Association for Computational Linguistics.

Qixun Wang, Yifei Wang, Hong Zhu, and Yisen Wang. 2022. Improving out-of-distribution generalization by adversarial training with structured priors. *Advances in Neural Information Processing Systems*, 35:27140–27152.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Zhaoyang Wang, Zhiyue Liu, Xiaopeng Zheng, Qinliang Su, and Jiahai Wang. 2023c. RMLM: A flexible defense framework for proactively mitigating word-level adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2757–2774, Toronto, Canada. Association for Computational Linguistics.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng, and Xuanjing Huang. 2023. Certified robustness to text adversarial attacks by randomized [MASK]. *Computational Linguistics*, 49(2):395–427.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

A Appendix

A.1 Datasets

HC3 (Guo et al., 2023): We conducted our in-domain experiments using the HC3 dataset, which is a compilation of human and ChatGPT answers from QA answers across four fields, including media, wiki, medicine, and finance. The dataset contains both English and Chinese data, but for our experiments, we focused on the English corpus. This corpus comprises 26,903 human texts and 58,546 ChatGPT texts. Following the methodology described in (Guo et al., 2023), we randomly partitioned the dataset, allocating 90% for training and 10% for testing. To evaluate the robustness of the AIGT detectors against adversarial perturbation attacks, we randomly selected 200 human-created samples and 200 AI-generated samples from the test set.

TruthfulQA (He et al., 2023): To evaluate the detector robustness in cross-domain scenarios, we utilized the TruthfulQA dataset. This dataset comprises human-created answers and AI-generated answers for 817 questions spanning 38 diverse fields, such as law, finance, health, and politics. It provides a suitable environment for assessing the cross-domain robustness of AI-generated text methods. In our experiments, we trained all the compared detectors on the HC3 train set and tested them on 817 human-created text and 817 ChatGPT-generated text of the TruthfulQA dataset. To evaluate the robustness of the detectors, we randomly selected 200 human-created samples and 200 ChatGPT-generated samples from TruthfulQA as test samples.

Ghostbuster (Verma et al., 2023): For evaluating the detector robustness in cross-genre scenarios, we employed the Ghostbuster dataset. This dataset consists of 3,000 parallel human-created articles and AI-generated articles from student essays, news articles, and creative writing. Similarly, we trained all the compared detectors on the HC3 train set, which compiles QA answers. We then assessed the detectors on articles from the Ghostbuster dataset. The robustness of detectors against adversarial attacks is evaluated by randomly selecting 200 human-created articles and 200 ChatGPT-generated articles from Ghostbuster dataset.

SeqXGPT-Bench (Wang et al., 2023a): To assess the detector robustness under mixed-source AI-generated texts, we utilized the SeqXGPT-Bench dataset. This dataset consists of parallel human-

created articles and AI-generated articles from various sources, including GPT-2, GPT-Neo, GPT-J, LLaMa, and GPT-3. It encompasses different domains such as news, social media, the web, scientific articles, and technical documents. For our experiments, we incorporated all 6,000 human-created texts and randomly selected one parallel AI-generated text from each of the five different AI sources, resulting in 6,000 parallel AI-generated texts from diverse sources. We allocated 90% of the data for training and reserved the remaining 10% as the test set. Again, for the adversarial perturbation evaluation, we randomly chose 200 human-created samples and 200 AI-generated samples from the test set.

A.2 Adversarial Perturbations

To evaluate the impact of adversarial perturbations on the AIGT detectors, we leverage three adversarial attack methods that encompass character-level and word-level substitution, deletion, and insertion.

PWWS (Ren et al., 2019), a widely utilized adversarial attack method, efficiently performs synonym substitution based on word saliency scores and maximum word-swap variance.

Deep-Word-Bug (Gao et al., 2018) incorporates random word-level substitution, swapping, deletion, and insertion, mimicking real-world human activities.

Pruthi (Pruthi et al., 2019) introduces adversarial perturbations by altering a small number of characters, resembling common typos. It encompasses character substitution, deletion, and insertion.

A.3 Baselines

To evaluate the performance of our proposed model, we compared it against several baseline models that are commonly used in AIGT detection. These baseline models include:

Log-Likelihood (Solaiman et al., 2019) employs the average token-level log probability generated by a language model as a feature. It trains a machine-learning classifier to determine whether a human or machine generates the text. A higher log probability indicates a higher likelihood that the text is AI-generated.

Log-Rank (Mitchell et al., 2023) utilizes the logarithm of the probability rank of each token generated by a decoder-only language model. A lower rank suggests a higher likelihood that the text is AI-generated.

Entropy (Gehrmann et al., 2019) computes the entropy value for each token based on the preceding text and then averages these scores to create the final feature.

GLTR (Gehrmann et al., 2019) constructs features based on the number of tokens in the top-10, top-100, top-1000, and top-1000+ ranks according to the token-level probability rank generated by the language models. The idea is that AI-generated tokens are more likely to belong to the head of the distribution during decoding.

SeqXGPT (Wang et al., 2023a) ensembles log probability scores from multiple language models and trains a CNN-based transformer model to detect the AI-generated sentences.

Bert (Devlin et al., 2019) is a widely used language model that demonstrates significant advancements in various NLP tasks, including many classification tasks.

Roberta (Liu et al., 2019) and **Deberta** (He et al., 2020) exhibit improved generalization ability compared to Bert, benefiting from additional training enhancements.

ChatGPT-Detector (Guo et al., 2023) is a RoBERTa-based detector tuned on the `HC3` dataset, which achieved state-of-the-art performance on ChatGPT text detection.

Additionally, we compared our method with recent robust methods that enhance classification robustness in other NLP tasks:

Flooding (Ishida et al., 2020) alleviates the model's overfitting by introducing an additional threshold into the training loss. This prevents the loss from decreasing further when it is already lower than the threshold, resulting in improved robustness.

RDrop (Wu et al., 2021) leverages dropout layers to generate similar text representations and then aligns the outputs to be the same. It reveals better generalization in several NLP tasks. In our experiments, we use a RoBERTa-base encoder as the base model.

RanMASK (Zeng et al., 2023) utilizes ensemble inference on several randomly masked text sequences copied from the input text. This approach enhances the model's robustness, particularly when encountering input word changes. In our experiments, we follow (Zeng et al., 2023) to use a RoBERTa-base encoder as the base model and set the mask percentage to 30%.

RMLM (Wang et al., 2023c) is an adversarial training method that adds extra adversarial samples

to the training data. It aims to defend the adversarial text attack by corrupting the adversarial text and then correcting the abnormal contexts. We follow the settings in their paper to train a BERT-based AIGT detector in our experiments.

A.4 Experiment Settings

We employed the RoBERTa-base encoder as the foundation of our model. To ensure a fair comparison, we selected the base versions of the pre-trained Bert, RoBERTa, and DeBERTa encoders used in the models under comparison. The detector training was conducted on 8 * 32GB NVIDIA V100 GPUs within the same environment. Subsequently, we evaluated the detectors under an adversarial perturbation attack on a single 32GB NVIDIA V100 GPU within the same environment.

Regarding the hyperparameters, we did not tune any of them in our experiment. For model training, we used a linear decay schedule with an initial learning rate of 1e-4. Consistent with (Guo et al., 2023), we trained all the compared detectors for 2 epochs on both the HC3 dataset and the SeqXGPT-Bench dataset. More detailed hyperparameter values can be found in Table 11.

To implement the adversarial attack methods, we utilized the open-source Textattack package (Morris et al., 2020). We implemented the compared models using their officially released code. For the ChatGPT-Detector, we directly utilized the model weight they released [1].

Hyperparameters	Value
Batch Size	16
Training Epochs	2
Optimizer	AdamW
Learning Rate	1e-4
d	768
d^z	512
α	2.0
β	0.5
λ_1	0.5
λ_2	0.01
λ_3	0.5

Table 11: Hyperparameters of our SCRN AIGT detector.

[1] https://huggingface.co/Hello-SimpleAI/chatgpt-detector-roberta

A.5 Threshold Analysis

In Section 4.2 - 4.5, we set the binary classification threshold to 0.5 to demonstrate the performance of AIGT detectors in general. However, the consequences of misclassifying human-created and AI-generated text in real-world scenarios may differ. Mislabeling human-created text as AI-generated text can have severe implications, such as affecting education exams and causing harm to innocent individuals. On the other hand, misclassifying AI-generated content as human-created can lead to the dissemination of harmful misinformation, resulting in public confusion or even social unrest.

To evaluate the robustness of AIGT detectors under different misclassification cost variations, we conducted experiments following the 1% false positive rate (FPR) setting proposed in (Hans et al., 2024; Krishna et al., 2024; Soto et al., 2024). Specifically, we evaluated the detectors under a fixed 1% FPR, where either AI-generated or human-created text is considered the positive sample. The training and testing were performed on `SeqXGPT-Bench`, which involves the generation of AI-generated text by mixed-source language models, better simulating the challenges encountered in real-world applications.

As shown in Table 12, when considering AI-generated text as positive and using the threshold under the 1% FPR setting, our SCRN detector achieves the highest AUA score, signifying superior robustness, while remaining competitive in terms of the OA score. Similarly, as illustrated in Table 13, when setting the threshold for 1% FPR with human-created text as positive samples, SCRN consistently demonstrates the best robustness, as indicated by the AUA score, and maintains competitive accuracy without attack, as measured by the OA score.

A.6 Inference Speed Comparison

Since the inference speed is an essential metric to evaluate the AIGT detector, we compared the inference speed of our proposed SCRN model with three other models: the RoBERTa backbone model, the previous state-of-the-art robust method RanMASK, and the adversarial training method RMLM.

Our SCRN model demonstrates highly effective inference speed when compared to the other AIGT models. As presented in Figure 3, in real-world inference service scenarios with a batch size of one, our model achieves an inference speed that is

Methods	AI → Human				Human → AI				Overall			
	OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑
Log-Likelihood	8.50	0.00	100.00	1150.41	98.50	98.50	**0.00**	**2116.83**	53.50	49.25	7.94	1633.62
Log-Rank	10.00	0.00	100.00	1212.20	99.50	**99.00**	0.50	2102.61	54.75	49.50	9.59	1657.41
Entropy	1.50	0.00	100.00	836.67	98.50	95.50	3.05	2106.64	50.00	47.75	4.50	1471.66
GLTR	17.50	0.00	100.00	1407.69	99.50	83.00	16.58	1985.72	58.50	41.50	29.06	1696.70
SeqXGPT	91.50	59.00	35.52	1787.49	**100.00**	89.50	10.50	2018.60	95.75	74.25	22.45	1903.04
BERT	73.00	0.00	100.00	1145.68	99.50	98.00	1.51	2097.26	86.25	49.00	43.19	1621.47
RoBERTa	91.00	62.00	31.87	1788.83	99.00	76.50	22.73	1879.16	95.00	69.25	27.11	1834.00
DeBERTa	91.50	46.50	49.18	1734.34	**100.00**	88.00	12.00	2020.35	95.75	67.25	29.77	1877.34
Flooding	**92.50**	57.00	38.38	1782.71	98.50	65.00	34.01	1770.68	95.50	61.00	36.13	1776.70
RDrop	**92.50**	68.00	26.49	1819.03	99.50	83.00	16.58	1914.17	**96.00**	75.50	21.35	1866.60
RMLM	84.50	68.00	19.53	1811.02	**100.00**	94.00	6.00	2024.73	92.25	81.00	12.20	1917.88
SCRN	89.00	**81.00**	**8.99**	**1972.56**	**100.00**	**99.00**	1.00	2101.37	94.50	**90.00**	**4.76**	**2036.96**

Table 12: AIGT detection results under fixed 1% false positive rate (FPR) considering **AI-generated text as positive samples**. Detectors are trained on the `Seqxgpt-Bench` training set, tested on randomly selected 200 human-created text and 200 AI-generated text from the `Seqxgpt-Bench` test set, and subjected to attacks using the PWWS method on the `Seqxgpt-Bench` attack set. The threshold is chosen to maintain an FPR of 1% across the entire test set. The best results are highlighted in bold.

Methods	AI → Human				Human → AI				Overall			
	OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑
Log-Likelihood	99.50	1.50	98.49	1521.43	0.50	0.00	100.00	504.00	50.00	0.75	98.50	1012.72
Log-Rank	99.50	1.00	98.99	1517.97	2.00	0.50	75.00	470.75	50.75	0.75	98.52	994.36
Entropy	98.50	0.50	99.49	1292.23	5.00	0.00	100.00	590.50	51.75	0.25	99.52	941.36
GLTR	99.50	1.50	98.49	1532.50	1.00	0.00	100.00	362.00	50.25	0.75	98.51	947.25
SeqXGPT	99.50	77.00	22.61	1934.58	**88.50**	54.50	38.42	1636.77	**94.00**	65.75	30.05	1785.68
BERT	99.00	11.00	88.89	1299.73	35.50	11.00	69.01	1981.18	67.25	11.00	83.64	1640.46
RoBERTa	99.00	75.50	23.74	1913.65	82.00	48.50	40.85	1640.80	90.50	62.00	31.49	1777.22
DeBERTa	99.00	71.00	28.28	1880.54	**88.50**	56.50	36.16	1648.58	93.75	63.75	32.00	1764.56
Flooding	99.50	69.50	30.15	1846.76	86.00	41.00	52.33	1449.56	92.75	55.25	40.43	1648.16
RDrop	**100.00**	74.50	25.50	1832.50	88.00	62.00	29.55	1742.41	**94.00**	68.25	27.39	1787.46
RMLM	99.00	80.50	18.69	1965.81	81.00	59.00	27.16	1678.57	90.00	69.75	22.50	1822.19
SCRN	98.50	**95.00**	**3.55**	**2013.52**	86.50	**72.00**	**16.76**	**1857.72**	92.50	**83.50**	**9.73**	**1935.62**

Table 13: AIGT detection results under fixed 1% false positive rate (FPR) regarding **human-created text as positive samples**. Detectors are trained on the `Seqxgpt-Bench` training set, tested on randomly selected 200 human-created text and 200 AI-generated text from the `Seqxgpt-Bench` test set, and subjected to attacks using the PWWS method on the `Seqxgpt-Bench` attack set. The threshold is chosen to maintain an FPR of 1% across the entire test set. The best results are highlighted in bold.

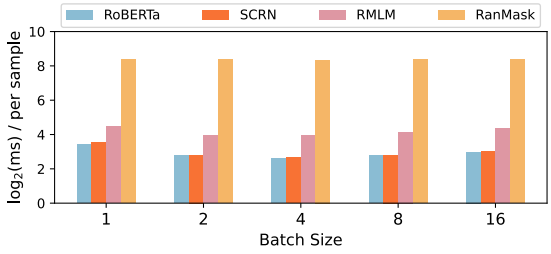

Figure 3: Inference time comparison of RoBERTa, RanMASK, RMLM, and SCRN on `HC3` test set. The experiments are conducted on a single 32GB NVIDIA-V100 GPU.

1.91× faster than RMLM. This speed advantage is primarily because RMLM involves result selection from a twice-forward process. Additionally, our model achieves an inference speed that is 28.5× faster than RanMASK, as RanMASK requires ensembling hundreds of results. When compared to the RoBERTa backbone detector, our model maintains 92.7% of RoBERTa's inference speed while surpassing up to 25.75 of accuracy under attack (AUA) achieved by RoBERTa as shown in Table 14.

For larger batch sizes, our SCRN model also demonstrates significantly faster inference speed compared to RMLM and RanMASK, with only a slight decrease in speed when compared to RoBERTa.

A.7 Inference Fluctuation Analysis

To assess the impact of inference fluctuation between two sub-model branches of SCRN, particularly due to randomness introduced by the reconstruction layer, we conducted a comparison of the detectors' accuracy under attack (AUA). Fig-

		OA ↑	AUA ↑	ASR ↓	ANQ ↑
In-domain	RoBERTa	**100.00**	74.67	25.33	5524.22
	RanMASK	**100.00**	79.25	20.75	5709.73
	RMLM	**100.00**	84.58	15.42	5631.18
	SCRN	**100.00**	**94.08**	**5.92**	**5716.17**
Corss-domain	RoBERTa	83.00	42.58	48.70	590.99
	RanMASK	85.00	45.50	46.47	462.53
	RMLM	89.75	51.58	42.53	535.79
	SCRN	**93.00**	**64.50**	**30.65**	**638.54**
Cross-genre	RoBERTa	**82.50**	31.00	62.42	17991.85
	RanMASK	77.00	30.25	60.71	15757.80
	RMLM	75.25	40.08	46.73	20728.78
	SCRN	82.25	**55.50**	**32.53**	**23179.41**
Mixed-source	RoBERTa	94.25	59.58	36.78	9644.89
	RanMASK	90.00	62.25	30.83	10127.65
	RMLM	91.25	72.17	20.91	11765.38
	SCRN	**95.33**	**85.33**	**10.50**	**13115.02**

Table 14: The robustness comparison of RoBERTa, RanMASK, RMLM, and SCRN. Results are average scores under three types of adversarial perturbation attacks including PWWS, Deep-Word-Bug, and Pruthi.

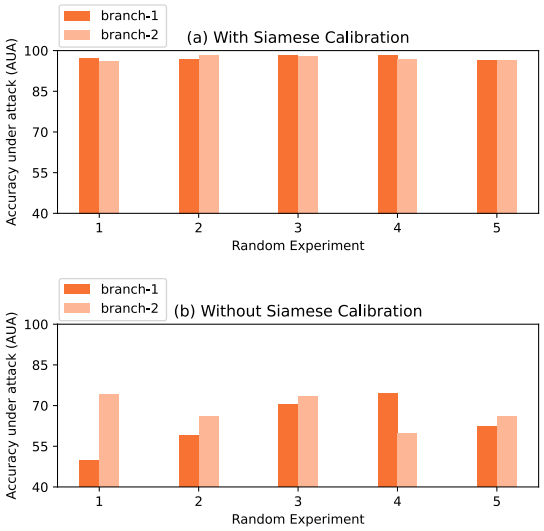

Figure 4: Inference fluctuation between two sub branches of SCRN on HC3 test set.

ure 4 presents the AUA fluctuation on HC3 dataset, considering five different random seeds. Notably, the AUA fluctuation remains below a 2.5 accuracy score across all scenarios. This observation indicates that the randomness introduced by the reconstruction process does not significantly affect the robustness of the SCRN detector. Thus, the gap between training and inference can be neglected.

This resilience can be attributed to the implementation of siamese calibration during the training of SCRN. The siamese calibration strategy aims to minimize the discrepancy in output distributions between the two branches of the model. Consequently, the classifier layer of SCRN is encouraged to prioritize high-level features over token-level features. This preference for high-level features enhances the model's robustness against token-level noise originating from the reconstruction process.

In contrast, when SCRN is trained without siamese calibration, the AUA fluctuation increases significantly, reaching up to a 14.75 accuracy score. This ablation experiment serves as further evidence of the effectiveness of siamese calibration in enhancing the robustness of the detector.

A.8 Case Study

As part of our research, we conducted a case study involving four cases from the Ghostbuster dataset, representing a cross-genre scenario. In this study, we compared our model with the previous state-of-the-art robust model, RMLM, which has been trained using adversarial techniques.

Figure 5 illustrates Case #1 and Case #2, which demonstrate word perturbations and character perturbations, respectively. While RMLM improves its robustness against adversarial perturbations by incorporating additional adversarial data during training, it fails to maintain robust detection in the cross-genre scenario. This observation highlights that simply augmenting the training set with adversarial data is not sufficient to effectively enhance the model's robustness against out-of-distribution samples (Wang et al., 2022). In contrast, our proposed SCRN model does not rely on extra prior information from the training set. As a result, SCRN successfully defends against cross-genre attacks, achieving superior robustness performance, as demonstrated in Table 16.

Moving on to Case #3 and Case #4, these cases explore the effects of word perturbations and character perturbations in human-created text, respectively. In comparison to AI-generated text, human-created text poses a greater challenge for attacks. It requires a higher proportion of manipulated words or characters to deceive the AIGT detector successfully.

A.9 Details of Cross-domain AIGT Detection under Adversarial Perturbations

Table 15 shows the robustness performance of all compared AIGT detectors on TruthfulQA dataset under three adversarial attack methods. The results consistently demonstrate the superior robustness of SCRN in cross-domain AIGT detection.

Figure 5: Cases from the Ghostbuster dataset are depicted in the figure. Case #1 and #2 represent AI-generated samples, whereas Case #3 and #4 are human-created samples. In cross-genre scenarios, RMLM fails to defend against adversarial text perturbations, whereas our SCRN demonstrates superior robustness. These cases highlight successful attacks on RMLM, while all adversarial attacks on these texts are unsuccessful against SCRN. Perturbed words or characters are [[**highlighted**]], while unchanged text is omitted for clarity.

A.10 Details of Cross-genre AIGT Detection under Adversarial Perturbations

Table 16 shows the robustness performance of all compared AIGT detectors on Ghostbuster dataset under three adversarial attack methods. The results consistently demonstrate the superior robustness of SCRN in cross-genre AIGT detection.

A.11 Details of Mixed-source AIGT Detection under Adversarial Perturbations

Table 17 shows the robustness performance of all compared AIGT detectors on SeqXGPT-Bench dataset under three adversarial attack methods. The results consistently demonstrate the superior robustness of SCRN in mixed-source AIGT detection.

A.12 AI Assistant Statement

Following the ACL 2023 Policy on AI Writing Assistance, we use AI assistant purely for the language of the paper, containing spell-checking, grammar-checking, and polishing our original content without suggesting new content. We affirm that all words refined by the AI assistant have been carefully reviewed and either rechecked or modified by us.

	Methods	AI → Human				Human → AI				Overall			
		OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑
PWWS	Log-Likelihood	67.00	0.00	100.00	381.32	97.00	97.00	0.00	89.96	82.00	48.50	40.85	235.64
	Log-Rank	72.00	0.00	100.00	374.13	95.50	95.00	0.52	89.63	83.75	47.50	43.28	231.88
	Entropy	43.50	0.00	100.00	433.91	98.00	80.00	18.37	83.96	70.75	40.00	43.46	258.94
	GLTR	66.50	0.00	100.00	376.47	88.50	56.50	36.16	79.80	77.50	28.25	63.55	228.14
	SeqXGPT	93.00	4.00	95.70	397.55	98.50	94.00	4.56	97.64	95.75	49.00	48.83	247.60
	BERT	80.00	0.00	100.00	384.23	99.50	99.50	0.00	89.24	89.75	49.75	44.57	236.74
	RoBERTa	90.50	6.50	92.82	410.38	75.50	74.00	1.99	99.81	83.00	40.25	51.51	255.10
	DeBERTa	91.50	1.00	98.91	381.80	98.00	97.50	0.51	88.78	94.75	49.25	48.02	235.29
	ChatGPT-Detector	96.00	1.00	98.96	364.00	98.50	88.50	10.15	85.93	97.25	44.75	53.98	224.96
	Flooding	90.00	0.00	100.00	387.93	78.00	74.50	4.49	101.59	84.00	37.25	55.65	244.76
	RDrop	89.50	6.00	93.30	429.06	89.00	73.50	17.42	91.10	89.25	39.75	55.46	260.08
	RanMASK	89.00	1.00	98.88	406.12	81.00	78.00	3.70	98.26	85.00	39.50	53.53	252.19
	RMLM	81.50	5.50	93.25	426.98	98.00	98.00	0.00	91.56	89.75	51.75	42.34	259.27
	SCRN	86.50	40.50	53.18	551.20	99.50	99.50	0.00	89.24	93.00	70.00	24.73	320.22
Deep-Word-Bug	Log-Likelihood	67.00	0.00	100.00	47.54	97.00	97.00	0.00	27.45	82.00	48.50	40.85	37.49
	Log-Rank	72.00	0.00	100.00	47.07	95.50	95.50	0.00	27.39	83.75	47.75	42.99	37.23
	Entropy	43.50	0.00	100.00	54.54	98.00	95.00	3.06	26.86	70.75	47.50	32.86	40.70
	GLTR	66.50	0.00	100.00	48.53	88.50	84.00	5.08	25.86	77.50	42.00	45.81	37.20
	SeqXGPT	93.00	0.50	99.46	56.35	98.50	98.50	0.00	27.13	95.75	49.50	48.30	41.74
	BERT	80.00	0.00	100.00	52.19	99.50	99.50	0.00	27.51	89.75	49.75	44.57	39.85
	RoBERTa	90.50	4.50	95.03	67.50	75.50	75.00	0.66	30.64	83.00	39.75	52.11	49.07
	DeBERTa	91.50	2.00	97.81	53.75	98.00	98.00	0.00	27.33	94.75	50.00	47.23	40.54
	ChatGPT-Detector	96.00	0.00	100.00	55.81	98.50	97.50	1.02	27.38	97.25	48.75	49.87	41.60
	Flooding	90.00	1.50	98.33	57.78	78.00	76.50	1.92	31.06	84.00	39.00	53.57	44.42
	RDrop	89.50	4.00	95.53	73.26	89.00	88.00	1.12	28.09	89.25	46.00	48.46	50.68
	RanMASK	89.00	3.00	96.63	68.10	81.00	75.00	7.41	27.84	85.00	39.00	54.12	47.97
	RMLM	81.50	2.50	96.93	56.86	98.00	98.00	0.00	28.35	89.75	50.25	44.01	42.60
	SCRN	86.50	17.00	80.35	107.34	99.50	99.50	0.00	27.50	93.00	58.25	37.37	67.42
Pruthi	Log-Likelihood	67.00	1.00	98.51	1849.10	97.00	97.00	0.00	151.68	82.00	49.00	40.24	1000.39
	Log-Rank	72.00	2.00	97.22	1938.40	95.50	95.50	0.00	151.69	83.75	48.75	41.79	1045.04
	Entropy	43.50	1.00	97.70	2329.26	98.00	93.50	4.59	150.90	70.75	47.25	33.22	1240.08
	GLTR	66.50	1.00	98.50	2025.11	88.50	82.00	7.34	147.90	77.50	41.50	46.45	1086.50
	SeqXGPT	93.00	3.50	96.24	2265.47	98.50	98.50	0.00	150.38	95.75	51.00	46.74	1207.92
	BERT	80.00	8.50	89.38	2733.19	99.50	99.50	0.00	151.35	89.75	54.00	39.83	1442.27
	RoBERTa	90.50	21.50	76.24	2767.34	75.50	74.00	1.99	170.28	83.00	47.75	42.47	1468.81
	DeBERTa	91.50	7.00	92.35	2387.38	98.00	97.00	1.02	150.17	94.75	52.00	45.12	1268.78
	ChatGPT-Detector	96.00	17.50	81.77	2693.03	98.50	91.00	7.61	151.59	97.25	54.25	44.22	1422.31
	Flooding	90.00	25.00	72.22	2873.47	78.00	76.50	1.92	172.08	84.00	50.75	39.58	1522.78
	RDrop	89.50	24.50	72.63	2813.49	89.00	87.00	2.25	155.13	89.25	55.75	37.54	1484.31
	RanMASK	89.00	36.00	59.55	2011.91	81.00	80.00	1.23	162.98	85.00	58.00	31.76	1087.44
	RMLM	81.50	7.50	90.80	2451.19	98.00	98.00	0.00	159.80	89.75	52.75	41.23	1305.50
	SCRN	86.50	31.00	64.16	2904.60	99.50	99.50	0.00	151.35	93.00	65.25	29.84	1527.98

Table 15: Results of cross-domain AIGT detection under different attack methods.

	Methods	AI → Human				Human → AI				Overall			
		OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑
PWWS	Log-Likelihood	62.00	0.00	100.00	2700.46	97.50	96.50	1.03	6077.26	79.75	48.25	39.50	4388.86
	Log-Rank	64.50	0.00	100.00	2734.98	97.50	95.50	2.05	6054.48	81.00	47.75	41.05	4394.73
	Entropy	77.50	0.00	100.00	2783.14	74.00	34.00	54.05	5352.47	75.75	17.00	77.56	4067.80
	GLTR	50.50	0.00	100.00	2696.80	97.50	67.50	30.77	5476.04	74.00	33.75	54.39	4086.42
	SeqXGPT	85.50	0.00	100.00	2712.97	88.00	65.50	25.57	5776.35	86.75	32.75	62.25	4244.66
	BERT	57.00	0.00	100.00	2692.61	95.50	75.00	21.47	5619.45	76.25	37.50	50.82	4156.03
	RoBERTa	82.00	0.00	100.00	2655.43	83.00	59.00	28.92	5522.05	82.50	29.50	64.24	4088.74
	DeBERTa	90.00	0.00	100.00	2763.66	77.50	53.50	30.97	5329.64	83.75	26.75	68.06	4046.65
	ChatGPT-Detector	58.50	0.00	100.00	2606.75	93.00	73.00	21.51	5827.88	75.75	36.50	51.82	4217.32
	Flooding	87.50	0.00	100.00	2733.18	82.50	58.00	29.70	5447.84	85.00	29.00	65.88	4090.51
	RDrop	95.00	10.00	89.47	3155.59	73.00	65.00	10.96	5973.84	84.00	37.50	55.36	4564.72
	RanMASK	67.00	2.00	97.01	2667.19	87.00	75.00	13.79	5433.82	77.00	38.50	50.00	4050.50
	RMLM	58.50	9.50	83.76	3397.99	92.00	72.50	21.20	5440.61	75.25	41.00	45.51	4419.30
	SCRN	94.50	71.00	24.87	4419.16	70.50	54.50	22.70	5725.79	82.50	62.75	23.94	5072.48
Deep-Word-Bug	Log-Likelihood	62.00	0.00	100.00	352.02	97.50	97.50	0.00	1726.71	79.75	48.75	38.87	1039.36
	Log-Rank	64.50	0.00	100.00	354.59	97.50	97.00	0.51	1726.94	81.00	48.50	40.12	1040.77
	Entropy	77.50	0.00	100.00	351.23	74.00	67.00	9.46	1613.55	75.75	33.50	55.78	982.39
	GLTR	50.50	0.00	100.00	354.47	97.50	93.50	4.10	1666.85	74.00	46.75	36.82	1010.66
	SeqXGPT	85.50	0.00	100.00	353.94	88.00	88.00	0.00	1724.10	86.75	44.00	49.28	1039.02
	BERT	57.00	0.00	100.00	368.96	95.50	90.00	5.76	1675.84	76.25	45.00	40.98	1022.40
	RoBERTa	82.00	0.50	99.39	351.32	83.00	71.00	14.46	1464.67	82.50	35.75	56.67	908.00
	DeBERTa	90.00	1.50	98.33	353.25	77.50	64.50	16.77	1316.41	83.75	33.00	60.60	834.83
	ChatGPT-Detector	58.50	0.50	99.15	346.69	93.00	87.00	6.45	1671.25	75.75	43.75	42.24	1008.97
	Flooding	87.50	4.00	95.43	359.58	82.50	68.00	17.58	1610.94	85.00	36.00	57.65	985.26
	RDrop	95.00	16.50	82.63	514.99	73.00	62.00	15.07	1597.97	84.00	39.25	53.27	1056.48
	RanMASK	67.00	6.00	91.04	445.55	87.00	60.00	31.03	1353.78	77.00	33.00	57.14	899.66
	RMLM	58.50	3.00	94.87	354.36	92.00	66.00	28.26	1492.68	75.25	34.50	54.15	923.52
	SCRN	94.50	50.00	47.09	1072.79	70.50	58.00	17.73	1535.02	82.50	54.00	34.55	1303.90
Pruthi	Log-Likelihood	62.00	0.00	100.00	48544.98	97.50	97.50	0.00	74465.43	79.75	48.75	38.87	61505.20
	Log-Rank	64.50	0.50	99.22	50907.28	97.50	95.50	2.05	74438.57	81.00	48.00	40.74	62672.92
	Entropy	77.50	0.00	100.00	37069.67	74.00	69.00	6.76	66872.37	75.75	34.50	54.46	51971.02
	GLTR	50.50	0.50	99.01	50694.88	97.50	95.00	2.56	74298.09	74.00	47.75	35.47	62496.48
	SeqXGPT	85.50	0.50	99.42	49312.26	88.00	88.00	0.00	73819.71	86.75	44.25	48.99	61565.98
	BERT	57.00	1.00	98.25	49171.64	95.50	87.50	8.37	64115.08	76.25	44.25	41.97	56643.36
	RoBERTa	82.00	1.50	98.17	37416.09	83.00	54.00	34.94	60541.50	82.50	27.75	66.36	48978.80
	DeBERTa	90.00	1.50	98.33	30985.48	77.50	42.00	45.81	43183.26	83.75	21.75	74.03	37084.37
	ChatGPT-Detector	58.50	0.00	100.00	24919.15	93.00	68.00	26.88	72614.00	75.75	34.00	55.12	48766.58
	Flooding	87.50	4.50	94.86	45652.25	82.50	54.50	33.94	51702.81	85.00	29.50	65.29	48677.53
	RDrop	95.00	24.00	74.74	64496.99	73.00	48.00	33.25	48799.05	84.00	36.00	57.14	56648.02
	RanMASK	67.00	8.00	88.06	45649.39	87.00	30.50	64.94	38997.11	77.00	19.25	75.00	42323.25
	RMLM	58.50	2.00	96.58	51019.47	92.00	87.50	4.89	62667.59	75.25	44.75	40.53	56843.53
	SCRN	94.50	51.50	45.50	86841.90	70.50	49.50	29.79	39481.41	82.50	50.50	38.79	63161.66

Table 16: Results of cross-genre AIGT detection under different attack methods.

	Methods	AI → Human				Human → AI				Overall			
		OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑	OA ↑	AUA ↑	ASR ↓	ANQ ↑
PWWS	Log-Likelihood	72.00	0.50	99.31	1281.86	62.00	53.50	13.71	1667.91	67.00	27.00	59.70	1474.88
	Log-Rank	73.50	0.50	99.32	1286.24	62.50	56.00	10.40	1697.20	68.00	28.25	58.46	1491.72
	Entropy	63.00	0.00	100.00	1239.29	55.50	27.50	50.45	1396.39	59.25	13.75	76.79	1317.84
	GLTR	76.50	0.00	100.00	1260.99	67.50	19.00	71.85	1285.64	72.00	9.50	86.81	1273.32
	SeqXGPT	96.50	65.00	32.64	1867.81	96.00	70.00	27.08	1893.98	96.25	67.50	29.87	1880.90
	BERT	90.50	1.00	98.90	1204.52	90.00	59.00	34.44	1815.44	90.25	30.00	66.76	1509.98
	RoBERTa	95.50	64.50	32.46	1840.19	93.00	62.50	32.80	1729.72	94.25	63.50	32.63	1784.96
	DeBERTa	95.50	54.50	42.93	1764.94	96.00	80.00	16.67	1940.47	95.75	67.25	29.77	1852.70
	Flooding	96.00	60.50	36.98	1800.01	95.50	53.00	44.50	1610.45	95.75	56.75	40.73	1705.23
	RDrop	96.50	69.00	28.50	1819.95	95.00	70.00	26.32	1815.62	95.75	69.50	27.42	1817.78
	RanMASK	94.00	60.00	36.17	1784.11	86.00	71.00	17.44	1715.72	90.00	65.50	27.22	1749.92
	RMLM	91.00	69.00	24.18	1879.96	91.50	78.00	14.75	1986.50	91.25	73.50	19.45	1933.23
	SCRN	95.00	87.00	8.42	1986.98	96.00	91.50	4.69	2099.91	95.50	89.25	6.54	2043.44
Deep-Word-Bug	Log-Likelihood	72.00	0.50	99.31	162.17	62.00	60.50	2.42	535.48	67.00	30.50	54.48	348.82
	Log-Rank	73.50	0.50	99.32	163.11	62.50	60.00	4.00	539.27	68.00	30.25	55.51	351.19
	Entropy	63.00	0.50	99.21	156.12	55.50	46.50	16.22	456.98	59.25	23.50	60.34	306.55
	GLTR	76.50	0.00	100.00	158.01	67.50	41.00	39.26	385.61	72.00	20.50	71.53	271.81
	SeqXGPT	96.50	2.00	97.93	171.14	96.00	68.50	28.65	547.31	96.25	35.25	63.38	359.23
	BERT	90.50	7.50	91.71	180.22	90.00	76.50	15.00	515.24	90.25	42.00	53.46	347.73
	RoBERTa	95.50	71.00	25.65	373.37	93.00	50.50	45.70	379.60	94.25	60.75	35.54	376.48
	DeBERTa	95.50	69.50	27.23	233.89	96.00	82.00	14.58	440.42	95.75	75.75	20.89	337.16
	Flooding	96.00	70.00	27.08	361.76	95.50	54.50	42.93	394.68	95.75	62.25	34.99	378.22
	RDrop	96.50	73.00	24.35	519.21	95.00	61.50	35.26	504.20	95.75	67.25	29.77	511.71
	RanMASK	94.00	57.50	38.83	464.99	86.00	75.00	12.79	622.30	90.00	66.25	26.39	543.64
	RMLM	91.00	73.50	19.23	428.07	91.50	73.00	20.21	541.52	91.25	73.25	19.73	484.79
	SCRN	95.00	83.00	12.63	586.14	96.00	91.50	4.69	640.43	95.50	87.25	8.64	613.28
Pruthi	Log-Likelihood	72.00	0.50	99.31	15511.22	62.00	60.00	3.23	33247.44	67.00	30.25	54.85	24379.33
	Log-Rank	73.50	1.00	98.64	17032.86	62.50	59.50	4.80	33279.58	68.00	30.25	55.51	25156.22
	Entropy	63.00	0.50	99.21	11240.34	55.50	36.00	35.14	27676.25	59.25	18.25	69.20	19458.29
	GLTR	76.50	0.00	100.00	12173.33	67.50	21.00	68.89	22492.95	72.00	10.50	85.42	17333.14
	SeqXGPT	96.50	1.00	98.96	18296.61	96.00	64.50	32.81	32981.87	96.25	32.75	65.97	25639.24
	BERT	90.50	1.00	98.90	14234.70	90.00	67.50	25.00	34276.58	90.25	34.25	62.05	24255.64
	RoBERTa	95.50	72.00	24.61	33905.64	93.00	37.00	60.22	19640.80	94.25	54.50	42.18	26773.22
	DeBERTa	95.50	72.50	24.08	34325.91	96.00	71.00	26.04	32058.96	95.75	71.75	25.07	33192.44
	Flooding	96.00	68.00	29.17	32915.51	95.50	30.00	68.59	20405.92	95.75	49.00	48.83	26660.72
	RDrop	96.50	70.50	26.94	32959.75	95.00	51.50	45.79	25669.62	95.75	61.00	36.29	29314.68
	RanMASK	94.00	52.00	44.68	30767.48	86.00	58.00	32.56	25411.29	90.00	55.00	38.89	28089.38
	RMLM	91.00	70.00	23.08	33542.23	91.50	69.50	24.04	32213.99	91.25	69.75	23.56	32878.11
	SCRN	94.50	80.50	14.81	36077.57	95.50	78.50	17.80	37299.11	95.00	79.50	16.32	36688.34

Table 17: Results of mixed-source AIGT detection under different attack methods.