Machine-Generated Text Localization

Zhongping Zhang   **Wenda Qin**   **Bryan A. Plummer**
Boston University
{zpzhang, wdqin, bplum}@bu.edu

Abstract

Machine-Generated Text (MGT) detection aims to identify a piece of text as machine or human written. Prior work has primarily formulated MGT detection as a binary classification task over an entire document, with limited work exploring cases where only part of a document is machine generated. This paper provides the first in-depth study of MGT that localizes the portions of a document that were machine generated. Thus, if a bad actor were to change a key portion of a news article to spread misinformation, whole document MGT detection may fail since the vast majority is human written, but our approach can succeed due to its granular approach. A key challenge in our MGT localization task is that short spans of text, *e.g.*, a single sentence, provides little information indicating if it is machine generated due to its short length. To address this, we leverage contextual information, where we predict whether multiple sentences are machine or human written at once. This enables our approach to identify changes in style or content to boost performance. A gain of 4-13% mean Average Precision (mAP) over prior work demonstrates the effectiveness of approach on five diverse datasets: GoodNews, VisualNews, WikiText, Essay, and WP. We release our implementation at this http URL.

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; Ouyang et al., 2022; Wang and Komatsuzaki, 2021; Gao et al., 2020; Du et al., 2022; Touvron et al., 2023) have led to significant advancements in many domains like conversational systems (OpenAI, 2023), social media data mining (Lyu et al., 2023), and medical image analysis (Nori et al., 2023), among others. Many ethical or factual problems can arise in text generation such as hallucination (Lin et al., 2022) or misuse for monetization (ad revenue through clicks) or propaganda (Zellers
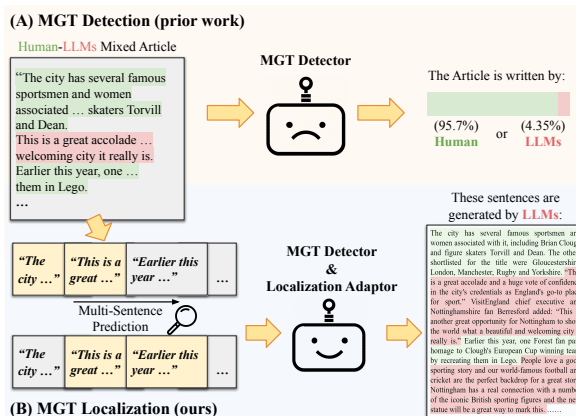

Figure 1: Prior work in machine-generated text detection (Mitchell et al., 2023; Su et al., 2023; Guo et al., 2023), shown in **(A)**, predicts a binary label indicating if an entire document or paragraph was machine or human generated. However, real-world articles may contain a mix of human-written and machine-generated sentences, which are challenging to detect when only a small part of the document was changed. To address this, we explore machine-generated text localization, shown in **(B)**, where we introduce a lightweight localization adaptor to perform sentence-level predictions within a text document. Our method predicts multiple sentences at once to address challenges caused by the text's short length.

et al., 2019). Machine-generated-text (MGT) detection can help defend against this misuse. However, as shown in Figure 1(A), prior work has primarily focused on whole document (*i.e.*, binary) classification as human or machine generated (*e.g.*, Tan et al., 2020; Mitchell et al., 2023; Verma et al., 2024; Guo et al., 2023; Su et al., 2023), but many applications may mix machine-generated and human-written text. For example, bad actors might use LLMs to manipulate certain sections of a news article to spread misinformation. Thus, whole document classification may fail since most text is human written. While Verma et al. (2024) did explore paragraph-level detection, this may still be too coarse to detect changes to single sentences.

8357

We also note that a concurrent work, Wang et al. (2023a), attempted to achieve sentence-level predictions. However, their approach primarily addresses articles where the initial segment is human-written and the second segment is AI-generated. In contrast, our paper considers a more challenging and general case in which multiple sections of an article can be generated by LLMs. A specific comparison is presented in Appendix B.2.

To bridge this gap, we introduce the first in-depth study on machine-generated text *localization*. As illustrated in Figure 1(B), our task's goal is to identify any machine generated sentences within a given article. A straightforward approach for our localization task would be simply employing sliding windows on top of existing detectors (Solaiman et al., 2019; Ouyang et al., 2022). While this enables us to adapt existing binary classification methods to our task (*e.g.*, OpenAI-Detector (Solaiman et al., 2019), DetectGPT (Wang et al., 2023b), and ChatGPT-Detector (Guo et al., 2023)), these detectors perform poorly on the inherently short length of sentences (*i.e.*, many of these models reported that reliable classification requires sentences to be longer than 50 tokens).

To address the aforementioned issues, we propose a lightweight Adaptor network for generated text Localization (AdaLoc). Our approach provides additional context by including multiple sentences at once, but then predicts whether each individual sentence is machine generated. This way the model has more information when making its predictions, but still produces dense labels. We find our approach can be further improved by aggregating overlapping predictions using a majority vote. Our experiments show our approach outperforms direct adaptations of state-of-the-art on MGT detection, training a model for MGT localization without context (*i.e.*, directly on the sentences it is trying to label), or aggregating overlapping predictions when a single label is produced over blocks of text.

In summary our contributions are:
- We provide the first in-depth study on machine-generated text localization[1]. This task bridges the gap between the current binary classification task and articles that contain a mix of human and machine-generated text.

- We introduce a data creation pipeline to generate articles consisting of both human-written and machine-generated texts. This approach can be used to automatically generate training and evaluation data for our MGT localization task.
- We identify a major challenge in machine-generated text localization arising from inaccurate judgements for short texts. To address this challenge, we use a majority vote strategy from overlapping predictions with our AdaLoc approach to provide dense labels over sentences in an article.
- The effectiveness of our proposed methods are validated on five diverse datasets (GoodNews, VisualNew, WikiText, Essay, and WP), with a 4∼13% mAP improvement.

2 Related Work

The importance of detecting machine-generated text has risen due to the risk of producing factual inaccuracies (Lin et al., 2022) and the potential for its use in misinformation, such as propaganda or monetization (Zhang et al., 2023). Existing detection methods can primarily be categorized into two types: metric-based methods and model-based methods. Metrics-base methods (Solaiman et al., 2019; Gehrmann et al., 2019; Mitchell et al., 2023; Su et al., 2023; Wang et al., 2023b) rely on extracting distinguishable features from text using the target language model. Specifically, Solaiman et al. (2019) apply log probability to identify whether a document is generated by LLMs or humans. Gehrmann et al. (2019) employ the absolute rank of each token as the evaluation metrics. Recent studies (Mitchell et al., 2023; Su et al., 2023; Bao et al., 2023) have shown that minor modifications to machine-generated text usually result in lower log probability under the model than the original text, a pattern not observed with human-written text. Thus, these methods introduce perturbations to the input text, measuring the discrepancy between the original and perturbed texts.

Model-based methods (Solaiman et al., 2019; Guo et al., 2023; Ippolito et al., 2020; Bhattacharjee et al., 2023) involve training specific classifiers on annotated corpora to classify input text directly. This kind of method is particular useful for detecting text generated by black-box or unknown models. For example, Solaiman et al. (2019) fine-tuned a RoBERTa (Liu et al., 2019) model based on outputs from GPT-series models. Guo et al. (2023)

---

[1] We note that there are tools like GPT-Zero (Tian and Cui, 2023) or Copyleaks (Copyleaks, 2023) capable of performing sentence-level analysis on machine-generated text. However, since these tools haven't released any papers about how they identify these sentences, we believe our claim is warranted.

developed their approach using the HC3 (Guo et al., 2023) dataset.

To improve the generalization capabilities of these detectors, Verma et al. (2024) extracted features from text using a series of language models and trained a classifier to categorize these features. However, all the methods we have discussed explored generated text detection at coarse scales (*i.e.*, the whole document or paragraph level). In contrast, our paper broadens the discussion to incorporate articles comprising both human-written and machine-generated content at a granular (sentence) level where prior work underperforms.

3 Machine-generated Text Localization

Given an article x containing sentences $S = \{s_1, ..., s_n\}$, Machine-Generated Text (MGT) localization aims at identifying specific sentences produced by LLMs. Unlike the MGT detection task that assigns a single label y for the whole document, our task predicts a sequence of labels $\{y_1, ..., y_n\}$, where each label y_i corresponds to an individual sentence s_i, providing a more precise indicator of machine-generated content in x.

A straightforward baseline to adapt existing methods (Mitchell et al., 2023; Su et al., 2023; Guo et al., 2023) to the localization task is predicting labels sentence by sentence[2] (sliding window). The major challenge here is that a single sentence often provides insufficient information to determine whether it is machine-generated due to its short length. To address this, we leverage the contextual information to improve performance, where our method predicts multiple sentences at once so that changes in style or content can be identified.

Specifically, Section 3.1 introduces our method for constructing manipulated articles, which serve as the training and evaluation data for our experiments. Section 3.2 present our methods to adapt established detectors to the MGT localization task. We first discuss a majority vote algorithm that predicts multiple sentences simultaneously to improve the single-sentence prediction. Given that this method assigns the same label to all sentences within a window, it requires a trade-off between the window size and the granularity of localization.

In order to enhance performance without sacrificing granularity, we propose a lightweight adaptor designed to predict multiple sentence at once and allocate a unique label to each. Figure 2 provides an overview of this method.

3.1 Data Preparation: Article Manipulation

As discussed in the Introduction, real-world articles might contain a mix of human-written and machine-generated text. To prepare such articles for our training and evaluation datasets, we use LLMs to produce sentences conditioned on the title and initial paragraphs of each article. Then, we substitute certain sections (*e.g.*, paragraphs or sentences) of the original article with these machine-generated sentences. Following Mitchell et al. (2023), we use a variety of language models for text generation[3], including GPT2-1.5B (Radford et al., 2019), GPTNeo-2.7B (Gao et al., 2020), GPTJ-6B (Wang and Komatsuzaki, 2021), OPT-2.7B (Zhang et al., 2022), and GPTNeoX-20B (Gao et al., 2020). During the generation process, we employ two sampling methods: top-k sampling with k set to 40, and top-p sampling with p ranging from 0.94 to 0.98. To maintain sentence integrity, we apply NLTK (Bird et al., 2009) for segmenting complete paragraphs into sentences, selecting only well-formed sentences for inclusion in the articles. Each article is combined with 1∼3 MGT segments, with segment lengths varying from 40 to 300 tokens.

3.2 Methods

Single-sentence Prediction. To achieve sentence-level predictions for MGT, one straightforward approach is to employ a sliding window technique, applying it to each sentence within an article using established methods (*e.g.*, Solaiman et al., 2019 or Guo et al., 2023). This strategy allows for the generation of a sequence of labels throughout the article, pinpointing specific sentences that are machine-generated. A major challenge with this approach is the short length of the input text. Consistent with previous studies (Solaiman et al., 2019; Mitchell et al., 2023; Verma et al., 2024), we find that MGT detectors produce reliable outcomes with inputs exceeding 50 tokens, while individual sentences typically fall short of this token count. To address this limitation, we propose expanding the window

---

[2]Although single-sentence prediction can be directly applied here, it cannot leverage context due to its single-input single-output nature. In contrast, MGT localization predicts a sequence labels $\{y_1, ..., y_n\}$ to identify AI-generated sentences given an article $S = \{s_1, ..., s_n\}$, improving prediction of each label y_i with context (*e.g.*, $s_{i-2}, s_{i-1}, s_{i+1}, s_{i+2}$).

[3]For Essay and WP (Verma et al., 2024) datasets, we directly combine the machine-generated text and human-written text to get such articles.
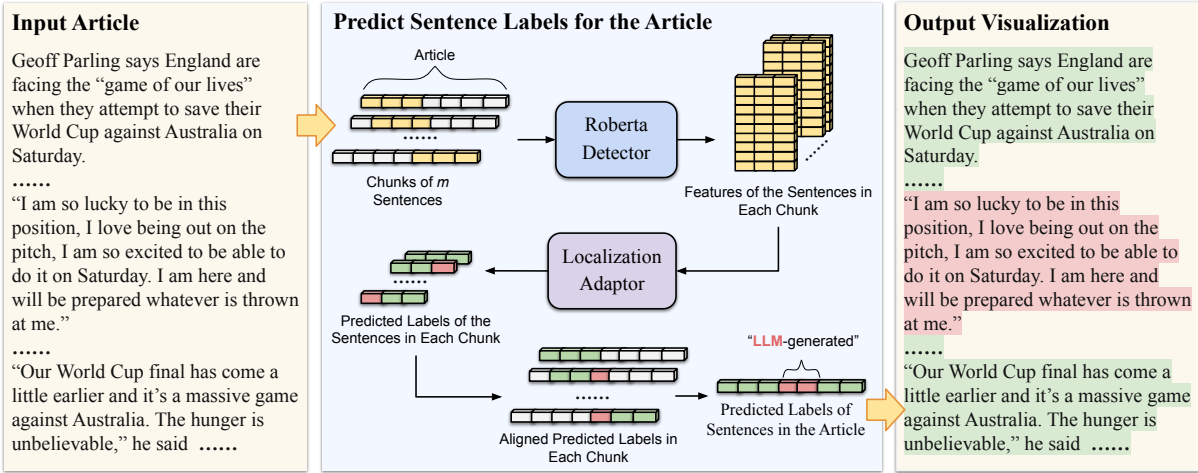
Figure 2: **Roberta+AdaLoc Overview.** Our method first divides an article into several chunks, each containing m sentences. We then employ existing MGT detection methods (*e.g.*, Solaiman et al., 2019 or Guo et al., 2023) to extract chunk features. The model parameters in this phase are fixed, eliminating the need for further training. To assign a specific label to each sentence, we introduce a lightweight localization adaptor, AdaLoc. AdaLoc consists of two fully connected layers, with the output from the final layer being an $m \times 1$ vector. This vector's elements represent the predicted labels for the sentences within the chunk. See Section 3.2 for detailed information.

size to incorporate multiple sentences at once, as detailed in the following paragraphs.

Multi-sentence Prediction with Majority Vote. In this approach, each window processes multiple sentences $\{s_1, ..., s_m\}$ as input, assigning the same label to all sentences within that window. Consequently, when the window step is set to 1, sentences within the same article are labeled m times. The final label for each sentence s_j is then determined based on these m labels, using a majority vote approach. Due to the uniform labeling within each window, this method necessitates balancing the window size against the localization granularity. As our experiments will show, this strategy improves performance compared to single-sentence predictions with an appropriate number of sentences per window, validating that the challenge of short text detection can be mitigated by increasing the window size.

Multi-sentence Prediction with Localization Adaptor. To boost the classification precision without reducing localization granularity, we further propose a lightweight localization adaptor, AdaLoc, capable of predicting multiple sentences simultaneously and assigning them corresponding labels. As shown in Figure 2, our method first divides the article into different chunks by NLTK (Bird et al., 2009), each comprising sentences $\{s_1, ..., s_m\}$. Leveraging the pretrained Roberta model (Solaiman et al., 2019), we obtain

chunk features with dimensions of 512×1024. The [CLS] token's (Liu et al., 2019) chunk features are extracted for the input to AdaLoc. AdaLoc consists of two Fully Connected (FC) layers configured as 1024-1024-m, incorporating dropout after the first FC layer. The output of AdaLoc is a $m \times 1$ vector $\{p_1, ..., p_m\}$, with each element denoting the label of a corresponding sentence within the chunk (window). Given m sentences, AdaLoc can predict 2^m potential binarization vectors. We apply binary cross entropy loss to finetune AdaLoc.

4 Experiments

4.1 Datasets & Metrics

GoodNews (Biten et al., 2019) is a news dataset that provides the URLs of New York Times articles from 2010 to 2018. After filtering out broken links and non-English articles, we randomly select 10,000, 1,000 and 1,000 articles for training, validation and test sets.

VisualNews (Liu et al., 2021) contains articles from four news sources: *Guardian*, *BBC*, *USA Today*, and *Washington Post*. Similarly, we randomly select 1,000 articles for evaluation. Another 1,000 articles are used to train logistic classifiers for metric-based methods like DetectGPT (Mitchell et al., 2023) and DetectLLM (Su et al., 2023).

WikiText (Stephen et al., 2017) contains 600/60/60 Wikipedia articles in training/validation/test sets,

Model	GPT-2	GPT-Neo	OPT	GPT-J	GPT-NeoX	**mAP**	**All**
Scale	-1.5B	-2.7B	-2.7B	-6B	-20B		
AP on GoodNews (Biten et al., 2019)							
Random	23.14	22.80	22.30	22.67	22.12	22.61	22.51
All 0/1	23.09	22.73	22.11	22.87	22.30	22.62	22.54
DetectGPT	48.91	48.87	46.40	49.87	46.69	48.15	47.53
DetectLLM	50.04	47.66	47.56	48.51	47.18	48.19	47.79
ChatGPT-D	32.51	31.35	30.94	30.10	28.85	30.75	30.64
Roberta-B	45.02	44.97	39.96	38.99	35.13	40.81	40.74
Roberta-L	57.24	58.05	49.38	47.41	41.32	50.67	50.85
Roberta-B+vote	60.92	61.31	55.56	53.63	48.51	55.99	56.24
Roberta-L+vote	71.03	72.03	64.37	62.28	55.39	65.02	65.50
Roberta-L+AdaLoc	**82.82**	**82.46**	**78.69**	**76.90**	**71.62**	**78.49**	**79.13**

Table 1: **Text Localization Results on GoodNews (Biten et al., 2019).** *vote* denotes multi-sentence prediction with majority vote, *AdaLoc* denotes multi-sentence prediction with our localization adaptor. For both methods, the window spans 3 sentences. To make a fair comparison, AdaLoc is finetuned only on GPT-2 generated articles, with the same procedure for Roberta-L. We observe that both *vote* and *AdaLoc* notably enhance the localization precision compared to single-sentence prediction. See Section 4.3 for detailed discussion.

respectively. We use the test set of WikiText directly for our evaluation.

Essay & WP (Verma et al., 2024) are designed for assessing AI-generated text detection in student essays and creative writings. Similar to the news articles, we randomly choose 1,000 human-authored documents from each dataset and blend them with AI-generated text (ChatGLM, ChatGPT, and GPT-4) for our analysis.

Metrics. Our experiments begin with the use of Average Precision (AP) to measure prediction accuracy for articles sampled from specific LLMs. We then compute the mean AP (mAP) based on documents generated by different LLMs. In addition, we aggregate predicted labels from all articles to calculate their collective AP (referred to as "All" in our comparisons). This metric allows us to evaluate a detector's performance across texts produced by various LLMs.

4.2 Baselines

Data Bias. We apply "Random" and "All 0/1" strategies to evaluate the data bias in our datasets.

DetectGPT (Mitchell et al., 2023) is a metric-based approach that introduces perturbations to the original text. This method is based on the intuition that LLM-derived text tends to be situated at the local optimum of the model's log probability function. Therefore, perturbations are likely to lower the log probability of machine-generated text, while the

effect on human-written text is more variable.

DetectLLM (Su et al., 2023) is a metric-based method which combines Log-Likelihood and Log-Rank (LRR) as its evaluation metric. For DetectGPT and DetectLLM, we trained a classifier for each domain's data distribution to determine the thresholds between machine-generated and human-written text.

ChatGPT-D (Guo et al., 2023) is proposed to detect texts generated by ChatGPT. This detector is trained on HC3 (Guo et al., 2023) dataset, which consists of 40k questions and their answers, written by both humans and ChatGPT.

Roberta-D (Solaiman et al., 2019) is a model trained on the output of GPT2, released by OpenAI. It can be generalized to outputs from other LLMs by fine-tuning with early stopping.

Roberta-MPU (Tian et al., 2024) is a framework designed to address the challenge of short-text detection. By incorporating a length-sensitive loss and a multiscale module, Roberta-MPU enhances the detection of short texts without compromising the performance on long texts.

GPT-zero (Tian and Cui, 2023) is an online tool for analyzing whether a piece of text is human-written or machine-generated. We apply it as an external, "blackbox" model to assess its performance in our localization task.

(A) Original Article	(B) Localization Results on Manipulated Article
Title: Three Drugs to Be Tested to Stave Off Alzheimer's **Body**: Maria C. Carrillo, vice president of medical and scientific relations at the Alzheimer's Association, said the results would come quickly. Within a few years, as researchers simultaneously compare the three approaches to stopping the disease, they should know which drug, if any, is going to work. "The association contributed $4.2 million to the study, more than twice as much as it has ever spent on a grant", Dr. Carrillo said. The announcement comes at a time of transition for Alzheimer's research. **......** The drugs were chosen from among 15 that drug companies offered, said the study's principal investigator, Dr. Randall Bateman of the Washington University School of Medicine in St. Louis. A committee assessed them, looking for drugs with the best evidence of effectiveness and the least likelihood of dangerous side effects. One concern is something called ARIA, for amyloid related imaging abnormality. People with the abnormality may have no signs that anything is wrong, but brain scans show what looks like a change in neural connections. **......** Researchers said they would face that issue when they come to it. "Right now we have to get treatments that work," said Dr. Rachelle S. Doody, director of the Alzheimer's Disease and Memory Disorders Center at the Baylor College of Medicine. "Then we can put pressure on to bring down the cost."	**Title**: Three Drugs to Be Tested to Stave Off Alzheimer's **Body**: Maria C. Carrillo, vice president of medical and scientific relations at the Alzheimer's Association, said the results would come quickly. Within a few years, as researchers simultaneously compare the three approaches to stopping the disease, they should know which drug, if any, is going to work. Carrillo said. "If there is a drug that works, we are going to be the ones to take it and test it", she said, "We are not going to be the ones to say no, But what about the people whose lives are most at risk?" The announcement comes at a time of transition for Alzheimer's research. **......** The drugs were chosen from among 15 that drug companies offered, said the study's principal investigator, Dr. Randall Bateman of the Washington University School of Medicine in St. Louis. Shouldn't a drug in development get tested in people who will be the most affected? The answer is no. The studies were not designed to test drugs in people who are at the highest risk for Alzheimer's disease. Because of that, their findings could have huge consequences for those in other developing countries. One concern is something called ARIA, for amyloid related imaging abnormality. People with the abnormality may have no signs that anything is wrong, but brain scans show what looks like a change in neural connections. **......** Researchers said they would face that issue when they come to it. "The study in the U.S., our conclusion is that we can't be confident in saying these drugs will work in the vast majority of the population," said Dr. William M. Foege, an associate professor of neurology and psychiatry at the University of California, San Francisco, "The study also showed that some of the drugs were unlikely to save lives. For example, the drug metformin, which can raise blood sugar, has so much side effects that most people with diabetes are put off by its side effects and don't use it at all." Then we can put pressure on to bring down the cost."

Figure 3: **A Qualitative Example on GoodNews.** We omit several human-written sections to fit the figure size. The machine-generated sentences are highlighted in light yellow and their original human-written sentences are highlighted in gray. Sentences localized by AdaLoc are marked by red color. We see that Roberta+AdaLoc effectively captures the manipulated segments in the article. See Section 4.3 for detailed discussion.

4.3 MGT Localization on GoodNews.

Quantitative Results. Table 1 presents the localization results of various models on GoodNews. We observe that both *vote* and *AdaLoc* boost the localization precision. For instance, Roberta-L+vote achieves a 15 mAP increase over single-sentence prediction methods. Incorporating AdaLoc yields an additional 13 mAP improvement over Roberta-L+vote.

We draw several conclusions from the Table. First, MGT localization appears more challenging than MGT detection. *E.g.*, while Roberta-Large can achieve over 80% accuracy in binary detection tasks (as per findings in Zhang et al., 2023 and Mitchell et al., 2023), it only achieves around 50 mAP in our localization task. Second, multi-sentence prediction methods outperform the single-sentence prediction strategy (*e.g.*, 40.8→55.9 in mAP of base size, 50.7→65.0 in mAP of large size), demonstrating that the challenge of detecting short texts can be alleviated by predicting multiple sentences together. Third, *AdaLoc* further boosts performance over *vote*, highlighting the importance of granularity in multi-sentence prediction.

In addition, our analysis reveals that the difficulty of MGT localization increases with the greater scale of LLMs. However, benefiting from direct access to the target language model, metric-based methods manage to maintain consistent per-

formance regardless of the model scale.

Qualitative Results. Figure 3 presents an article example from GoodNews. From the Figure, we see that the primary messages conveyed by the article can be substantially altered with just a few manipulated sentences (noting the contrast between the original sentences in gray and the machine-generated sentences in light yellow), emphasizing the importance of MGT localization. We observe that while sentences in boundary may occasionally be misidentified (*e.g.*, "Becaused of that, their findings .... countries." is a machine-generated sentence but was misidentified as human-written), AdaLoc can accurately localize the majority of the machine-generated text segments (marked by red). These findings demonstrate that a reliable MGT localization approach can help people defend against misinformation in manipulated articles.

Ablation Study on Window Size. As discussed in Section 3.2, our multi-sentence prediction algorithms need to find a balance between window size and granularity. Table 3 provides an ablation study on window size vs. the number of segments. In our data generation process, greater number of segments leads to shorter individual segment lengths. Specifically, Segs = 1, 2, and 3 corresponds to average segment lengths of 168.8, 84.2, and 57.7 tokens, respectively.

We observe that larger window sizes typically

Model	GPT-2	GPT-Neo	OPT	GPT-J	GPT-NeoX	mAP	All
Scale	-1.5B	-2.7B	-2.7B	-6B	-20B		
(A) AP on VisualNews (Liu et al., 2021)							
Random	16.70	16.95	16.53	17.13	16.80	16.83	16.73
All 0/1	16.91	16.71	16.73	17.36	16.71	16.88	16.81
DetectGPT	37.13	37.51	35.89	35.62	36.82	36.59	36.93
DetectLLM	38.69	37.73	39.91	38.57	38.26	38.63	38.38
ChatGPT-D	25.05	23.21	22.59	23.13	21.48	23.09	22.95
Roberta-B	36.93	36.87	33.27	31.33	27.50	33.18	33.01
Roberta-L	49.18	51.35	41.71	39.86	32.85	42.99	43.06
Roberta-B+vote	53.84	53.39	47.63	46.47	39.47	48.16	48.36
Roberta-L+vote	66.79	66.62	58.59	56.74	47.55	59.26	59.69
Roberta-L+AdaLoc	**78.40**	**78.29**	**72.37**	**70.96**	**64.46**	**72.90**	**73.35**
(B) AP on WikiText (Stephen et al., 2017)							
Random Guess	15.38	14.19	14.49	13.23	14.25	14.31	14.02
All 0/1	14.60	13.99	14.33	13.08	13.47	13.89	13.87
Roberta-B	36.63	32.31	30.90	23.80	21.18	28.96	29.00
Roberta-L	45.39	40.98	35.42	28.93	23.67	34.88	34.98
Roberta-B+vote	51.56	43.61	40.67	30.29	27.48	38.72	39.40
Roberta-L+vote	64.69	58.26	49.47	40.60	33.99	49.40	50.68
Roberta-L+AdaLoc	**74.55**	**70.89**	**66.83**	**57.53**	**54.70**	**64.90**	**66.03**

Table 2: **Zero-shot Localization Results on VisualNews and WikiText.** Despite being fine-tuned only on GoodNews articles, Roberta-L+AdaLoc boosts performance over Roberta-L+vote on both VisualNews and WikiText, achieving 13.7% and 15.5% mAP increases. These gains indicate that AdaLoc is able to identify LLM-generated sentences without overfitting to specific human-written styles in GoodNews. See Section 4.4 for detailed discussion.

Size	Segs=1	Segs=2	Segs=3	Avg.
$m=1$	55.89	57.19	58.62	57.23
$m=2$	74.03	70.12	**66.29**	70.15
$m=3$	78.40	**70.59**	64.11	**71.03**
$m=4$	**78.42**	66.49	59.85	68.25
$m=5$	76.46	61.57	55.99	64.67

Table 3: **Ablation Study of Window Size on Good-News.** m denotes the number of sentences in a sliding window, "Segs" denotes the number of machine-generated segments in an article. With our data generation method, greater number of segments results in shorter text length per segment. We observe that greater m leads to improved performance on segments of long texts, while reduced precision on segments with short texts. See Section 4.3 for detailed discussion.

Model	mAP	All
Roberta-L	50.67	50.85
Roberta-L+vote	65.02	65.50
Roberta-L+AdaLoc(skip)	67.59	67.63
Roberta-L+AdaLoc(middle)	70.54	70.63
Roberta-L+AdaLoc	**78.49**	**79.13**

Table 4: **Ablation Study of Vote Strategy on Good-News.** "skip" denotes that there is no overlapping between different chunks, *i.e.*, window step equals three sentences. "middle" means we leverage AdaLoc to predict whether the sentence in the middle is machine-generated. By default, AdaLoc is combined with the majority vote strategy. See Section 4.3 for discussion.

predict three sentences at once within a window.

Ablation Study on Vote Strategy. Table 4 provides ablation studies for the vote strategy within AdaLoc. We observe that AdaLoc, when combined with the majority vote strategy, achieves the best performance. Alternative strategies, such as "skip" and "middle" achieves lower perfor-

performs better on longer segments, with a reduced precision on shorter segments. m ranging from 2 to 4 correspond to the optimal performance for segment numbers ranging from 1 to 3. Based on the average values across different segment numbers, we set m to 3, *i.e.*, our *vote* and *AdaLoc* methods

Method	Essay (Verma et al., 2024)					WP (Verma et al., 2024)				
	GLM	GPT3.5	GPT3.5t	GPT4All	**mAP**	GLM	GPT3.5	GPT3.5t	GPT4All	**mAP**
Random	20.89	27.31	49.02	24.76	30.50	19.99	24.65	42.33	20.96	26.98
All 0/1	20.71	27.26	49.12	24.66	30.44	20.06	24.63	42.05	21.09	26.96
Roberta-L	42.27	33.10	47.70	38.07	40.28	56.87	29.18	32.35	36.76	38.80
Chat-D	64.47	54.94	68.43	55.50	60.84	53.53	43.17	51.48	48.75	49.23
Roberta-MPU	**70.26**	60.28	67.37	60.26	64.54	60.18	45.25	**60.81**	54.43	55.17
Roberta-L+vote	55.46	45.62	49.32	50.27	50.17	**70.22**	41.88	33.14	**62.05**	51.82
Chat-D+vote	62.16	59.15	75.17	58.87	63.84	57.63	44.79	58.09	54.15	53.67
Chat-D+AdaLoc	67.49	**68.73**	**78.53**	**61.04**	**68.95**	61.83	**51.33**	60.58	56.38	**57.53**

Table 5: **Text Localization Results on Essay and WP (Verma et al., 2024).** *AdaLoc* boosts the performance based on ChatGPT-D, demonstrating its generalization ability across different detectors in diverse domains. See Section 4.5 for discussion.

mance. Detailed ablation results are presented in Appendix B.4.

4.4 Zero-shot Experiments on VisualNews and WikiText

Experimental results in Section 4.3 on GoodNews show that Roberta+AdaLoc outperforms baselines when evaluated on in-domain data, where the training and test sets have similar data distributions. To verify our model is not simply overfitting to GoodNews, we perform zero-shot experiments on VisualNews and WikiText articles, as presented in Table 2. In zero-shot evaluations, Roberta-L+vote and Roberta-L+AdaLoc continue to boost base models in mAP (43.1→59.7→73.4 on VisualNews and 34.9→50.7→66.0 on WikiText). It illustrates that the improvements offered by *vote* and *AdaLoc* are due to their effectiveness in indentifying LLM-generated text, rather than recognizing the specific human-written styles found in GoodNews articles.

4.5 MGT Localization on Essay and WP

In previous sections, we primarily focused on MGT localization in long articles, such as news reports and Wikipedia. To evaluate our approach in diverse domains, we extend our discussion to student essays (Essay) and creative writing (WP) datasets (Verma et al., 2024). Following Verma et al. (2024), our experiments focuses on detecting text generated by ChatGLM (Du et al., 2022), ChatGPT3.5 (Ouyang et al., 2022), ChatGPT3.5-turbo (Ouyang et al., 2022), and GPT4All (Anand et al., 2023). Table 5 presents the localization results of different models. Since the training data of ChatGPT-D (Chat-D) (Guo et al., 2023) already includes a substantial number of ChatGPT-generated

Method	Precision	Recall
GPT-Zero	64.29	18.92
ChatGPT+AdaLoc	**84.15**	**40.47**

Table 6: **Comparison to GPT-Zero (Tian and Cui, 2023) on sentence-level analysis.** In our experiments, we observe that GPT-Zero often classifies manipulated articles as human-written, primarily because these articles contain significant portions of human-written text. As a results, GPT-Zero tends to label all sentences in these documents as human-written and fails to identify machine-generated sentences. See Section 4.6 for detailed discussion.

content, we utilize Chat-D as our backbone to extract chunk features.

The Table illustrates that our methods, *vote* and *AdaLoc* continue to improve the base model's localization performance, verifying that our method's adaptability to various detectors. To validate the effectiveness of context, we further compare Chat-D+AdaLoc to Roberta-MPU (Tian et al., 2024), a concurrent method for single sentence prediction. From the results, Chat-D+AdaLoc outperforms Roberta-MPU by 2-4 mAP. Roberta-MPU achieves comparable or slightly better results than Chat-D+vote (1-2 mAP boost). We note that Roberta-MPU requires training/finetuning the Roberta architecture on corpora with various lengths of text. In contrast, our Chat-D+vote achieves comparable results to Roberta-MPU by context and does not require any additional training.

4.6 Comparison to GPTZero

In our experiments, we note that the online tool GPTZero (Tian and Cui, 2023) can provide an

	Method	ChatGLM	GPT3.5	GPT3.5-t	GPT4All	mAP
(A)	Roberta-L+vote	55.46	45.62	49.32	50.27	50.17
	ChatGPT-D+vote	62.16	59.15	75.17	58.87	63.84
	Roberta-L+AdaLoc (VisualNews)	63.05	52.29	51.02	59.65	56.50
	ChatGPT-D+AdaLoc	**67.49**	**68.73**	**78.53**	**61.04**	**68.95**
(B)	Roberta-L+vote	70.22	41.88	33.14	62.05	51.82
	ChatGPT-D+vote	57.63	44.79	58.09	54.15	53.67
	Roberta-L+AdaLoc (VisualNews)	**75.97**	50.59	34.48	**65.11**	56.54
	ChatGPT-D+AdaLoc	61.83	**51.33**	**60.58**	56.38	**57.53**

Table 7: **Out-of-domain Evaluation on Various Document Categories.** Roberta-L+AdaLoc (VisualNews) is finetuned only on VisualNews articles and then evaluated on **(A)** Essay and **(B)** WP. We observe that Roberta-L+AdaLoc(VisualNews) achieves 5∼6 mAP boost compared to Roberta-L+vote, validating the generalization ability of AdaLoc. See Section 4.7 for discussion.

analysis about which sentences are likely to be AI-generated, which can be used for our localization task. Therefore, we include an additional comparison to GPT-Zero in this Section. Specifically, we randomly select 30 ChatGPT-manipulated articles from the Essay dataset and apply GPT-Zero to analyze which sentences are machine-generated.

Table 6 reports the comparison between GPTZero and ChatGPT-D+AdaLoc. We observe that GPTZero struggles to retrieve machine-generated sentences in manipulated articles. This is because these articles contain only a few portions generated by LLMs, and are likely to be classified as human-written by GPTZero. In these cases, all sentences in such documents are categorized to human-written, leading to low recall scores. We provide a specific example in Appendix B.3. We also tested 30 GPT2-manipulated articles on GoodNews, finding that GPTZero encounters the domain shift issue. That is, all GPT2-manipulated articles are classified as human-written.

4.7 Out-of-domain Evaluation

In Section 4.4, we validated through zero-shot experiments that our method is capable of recognizing machine-generated sentences in news articles from various sources. To further evaluate AdaLoc's generalization across different types of documents (*e.g.*, news articles, study essays, and creative writings), we conducted out-of-domain experiments on Essay and WP. Specifically, we first fine-tuned Roberta-L+AdaLoc using VisualNews articles, and then directly applied Roberta-L+AdaLoc (VisualNews) to Essay and WP. Table 7 presents the experiment results. From the table, we observe that Roberta-L+AdaLoc (VisualNews) outperforms

Roberta-L+vote, achieving a 5∼6 mAP improvement. Since our adaptor is fine-tuned only on news articles, the improvements indicate that AdaLoc has a strong generalization ability, even across different document categories including news articles, student essays, and creative writings. This conclusion is also consistent with our zero-shot experiment results in Section 4.4.

5 Conclusion

In this paper, we conduct a comprehensive study of MGT localization, aiming at recognizing AI-generated sentences within a document. We identify a major challenge in MGT localization as the short spans of text at the sentence level, *i.e.*, single sentence may not provide sufficient information for distinguishing machine-generated content. To address this, we propose our methods, *vote* and *AdaLoc*, to predict multiple sentences together, allowing changes in style or content to boost performance. Our methods are evaluated on five diverse datasets (GoodNews, VisualNews, WikiText, Essay, and WP), achieving a gain of 4∼13% mAP over baselines. The improvements across various datasets and detectors demonstrate the effectiveness and generalization of our method in MGT localization.

Limitations

In this paper, we highlight short text detection as a primary challenge in MGT localization and introduce our methods, *vote* and *AdaLoc*, to enhance the efficacy of existing detectors. Despite notable improvements across diverse datasets, our findings indicate substantial potential for further enhancement localization metrics, particularly in identifying sentences produced by more advanced language models. For example, ChatGPT+AdaLoc achieves only 69% and 57% mAP on Essay and WP, respectively. Roberta-L+AdaLoc obtains 65∼78% AP for texts generated by GPT-NeoX. Therefore, detection of short texts remains a challenge to be further explored.

Another challenge in our experiments is the domain-shift issue, where detectors optimized for one domain often experience varying degrees of performance degradation on out-of-domain data. For example, in our experiments, detectors all achieve lower scores on VisualNews and WikiText compared to GoodNews. Thus, enhancing model's generalization across different domains, such as combining GhostBuster (Verma et al., 2024) with our method, could be a potential direction for further work.

In addition, our method is specifically developed to localize the text that is generated by machines directly. Instances where human-written text is paraphrased by LLMs or vice versa are not considered in our study. Therefore, exploring approaches to identify paraphrased content within articles, such as Krishna et al. (2024), represents another area for further work.

Ethics Statement

In our study, we propose MGT localization methods (*e.g.*, Roberta-L+AdaLoc or ChatGPT-D+AdaLoc) to identify LLM-generated sentences within text documents, which can be helpful in defending against misinformation spread by LLMs. However, like all other detectors, our system will not produce 100% accurate predictions, especially when detecting texts from models unseen during training, as well as text domains that are far from the training corpus. Therefore, we strongly discourage incorporating our methods into automatic detection systems without human supervision, such as plagiarism detection or other situations involving suspected use of LLM-generated text. A more suitable application case would be using our methods under human supervision, detecting misinformation generated by LLMs in articles or social media content. We also recognize that bad actors could manipulate articles and spread misinformation according to the data preparation pipeline presented in our paper. Thus, we aim for our paper to highlight the need for building tools like Roberta+AdaLoc to identify and localize manipulated content in such articles.

References

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https://github.com/nomic-ai/gpt4all.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. 2023. Conda: Contrastive domain adaptation for ai-generated text detection. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 598–610.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12466–12475.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Copyleaks. 2023. https://copyleaks.com.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 320–335.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M 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.

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*.

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.

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.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visualnews: Benchmark and challenges in entity-aware image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771.

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*.

Hanjia Lyu, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo. 2023. Gpt-4v (ision) as a social media analysis engine. *arXiv preprint arXiv:2311.07547*.

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*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:13.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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*.

Merity Stephen, Xiong Caiming, Bradbury James, and Richard Socher. 2017. Pointer sentinel mixture models. *Proceedings of ICLR*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Reuben Tan, Bryan A. Plummer, and Kate Saenko. 2020. Detecting cross-modal inconsistency to defend against neural fake news. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Edward Tian and Alexander Cui. 2023. Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, Qinghua Zhang, Ruifeng Li, Chao Xu, and Yunhe Wang. 2024. Multiscale positive-unlabeled detection of AI-generated texts. In *The Twelfth International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

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.

Rongsheng Wang, Qi Li, and Sihong Xie. 2023b. Detectgpt-sc: Improving detection of text generated by large language models through self-consistency with masked predictions. *arXiv preprint arXiv:2310.14479*.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhongping Zhang, Yiwen Gu, and Bryan A. Plummer. 2023. Show, write, and retrieve: Entity-aware article generation and retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

A   Implementation Details

In our experiments, we utilized the open-source toolkit, MGTBench (He et al., 2023), to evaluate various baselines, such as DetectGPT and DetectLLM. Our model is primarily developed based on Pytorch (Paszke et al., 2019) and Transformers (Wolf et al., 2020) libraries. We configured our detector to a maximum sequence length of 512 tokens. For Roberta+AdaLoc and ChatGPT-D+AdaLoc, we adopted a batch size of 16 and set the learning rate to 1×10^{-5}. Though increasing the number of training epochs can result in better performance, we finetuned AdaLoc for three epochs

with an early stopping mechanism to prevent overfitting to specific data domains, following Verma et al. (2024). Our experiments were conducted on NVIDIA RTX-A6000 or A40 GPUs, fitting a 48 GB GPU memory requirement. It takes approximate one hour to finetune AdaLoc on chunks extracted from 10,000 manipulated articles.

B   Additional Results

B.1   Visualization of Manipulated Articles

As outlined in Section 3.1, we use language models to produce sentences and blend these generated sentences with human-written text to create our training and evaluation data. Figure 4 provides a specific example of this process as a supplementary example to the main paper. In this example, we replaced two segments of human-written text with content generated by GPT-J (Wang and Komatsuzaki, 2021), highlighted in light yellow and pink, respectively.

B.2   Comparison to SeqXGPT

To demonstrate the differences between our MGT localiztion task and sentence prediction in SeqXGPT (Wang et al., 2023a), we provide a specific example in Figure 5. The Figure illustrates that SeqXGPT follows an assumption that articles are structured with an initial human-written segment followed by an AI-generated segment. In contrast, our task involves articles containing multiple machine-generated sections, aligning more closely with real-world application scenarios. In addition, our synthetic data provides more abundant annotations, including sentence-level labels, LLM sampling strategies, and the number of machine-generated segments.

B.3   Article Assessment Interface of GPT-Zero

A screenshot of an article evaluated by GPT-Zero is provided in Figure 6, supplementing our main paper. From the Figure, we see that GPT-Zero incorrectly identifies the manipulated article as human-written, labeling all sentences within as human-written, which results in low recall scores. This instance supports our discussion in Section 4.6.

B.4   Ablation Study

In Section 4.3, we mainly ablate the window size, *i.e.*, predicting how many sentences at once within a window. Table 8 provides additional ablation

studies of AdaLoc. We see that AdaLoc, when combined with the majority vote strategy, achieves the best performance. Alternative strategies, such as "skip" and "middle" achieves lower performance.

(A) Human-written Article

Title: Nottingham named as 'Home of English Sport'.
Publish date: 10-23-2015
Domain: www.bbc.com
Body:
Nottingham has been named as England's official Home of Sport following a campaign by tourism body VisitEngland. The campaign, which included an online poll, was run in a bid to find the country's top sporting destination.

The city has several famous sportsmen and women associated with it, including Brian Clough and figure skaters Torvill and Dean. The others shortlisted for the title were Gloucestershire, London, Manchester, Rugby and Yorkshire.

Jennifer Spencer, chief executive of Experience Nottinghamshire, said the accolade has the potential to bring more visitors and help the city bid for major sporting events. Brian Clough, the former Nottingham Forest manager, had a huge impact on the city and was honoured with a statue in the city centre.

Earlier this year, one Forest fan paid homage to Clough's European Cup winning team by recreating them in Lego.

A film, I Believe in Miracles, about the players was also premiered at the City Ground, with director Johnny Owen describing them as \"gods\". The filmmaker tweeted about his delight at the honour and so did his partner, the Nottingham-born This is England star Vicky McClure.

(B) Article Manipulated by GPT-J

Title: Nottingham named as 'Home of English Sport'.
Publish date: 10-23-2015
Domain: www.bbc.com
Body:
Nottingham has been named as England's official Home of Sport following a campaign by tourism body VisitEngland.The campaign, which included an online poll, was run in a bid to find the country's top sporting destination.

The city has several famous sportsmen and women associated with it, including Brian Clough and figure skaters Torvill and Dean. The others shortlisted for the title were Gloucestershire, London, Manchester, Rugby and Yorkshire.

"This is a great accolade and a huge vote of confidence in the city's credentials as England's go-to place for sport." VisitEngland chief executive and Nottinghamshire fan Berresford added: "This is another great opportunity for Nottingham to show the world what a beautiful and welcoming city it really is."

Earlier this year, one Forest fan paid homage to Clough's European Cup winning team by recreating them in Lego.

People love a good sporting story and our world-famous football and cricket are the perfect backdrop for a great story. Nottingham has a real connection with a number of the iconic British sporting figures and the new statue will be a great way to mark this.

Figure 4: **A synthesized article consists of both human-written and machine-generated texts.** (A): presents the original human-written article from VisualNews (Liu et al., 2021); (B): an synthesized article manipulated by GPT-J (Wang and Komatsuzaki, 2021). In this example, we replaced two paragraphs of the original article with machine-generated content, highlighted in yellow and pink. See Appendix B.1 for discussion.

(A) A Sample from SeqXGPT

text: high - salt has been shown to play a role in the pathogenesis of autoimmune disease. in this study , we investigated the effect of high - salt on the production of inflammatory mediators by arpe-19 cells and the possible mechanisms involved. **…..** In this study, arpe-19 cells have been treated for 24 hours with lps at different concentrations which resulted to significantly different results. lps-treated arpe-19 cells grew as compared to untreated cells in the range of 40.3 - 57.3% of the initial culture volume. This increase in proliferation indicates that lps acts via the activation of arpe-19 cells immune mechanism.\n\nLps has also been recently used for the anti-tumour effect in cancer cell line. in this study, we found that Lps can bind to EBOV-1 and -2 and promote the cell death,
prompt_len: 347,
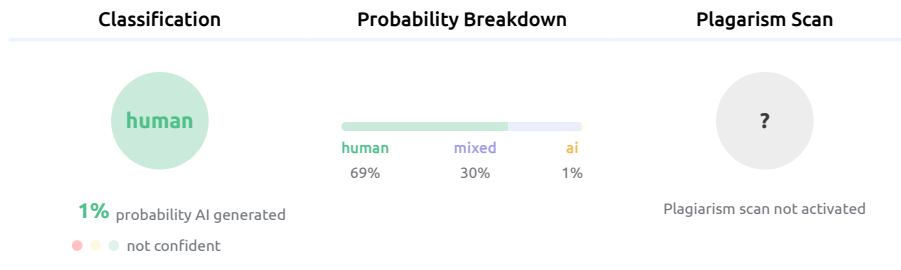**label**: gpt2

(B) A Sample from Our Synthetic Data

original_article:
title: Ace Favors Fewer Starts to Protect Pitchers' Arms .
Perhaps a six-man rotation could have helped alleviate the stress, not only on Tanaka, but all the other pitchers in the rotation, and all across baseball.
……
He is scheduled to make his 19th start Wednesday, and if he beats the Yankees, he will have won at least 10 games in his first three seasons.
Over all, he is 38-23 with a 3.24 E.R.A., averaging 11.2 strikeouts per nine innings. Last year, he led the major leagues with 277 strikeouts.
manipulated_article:
title: Ace Favors Fewer Starts to Protect Pitchers' Arms .
Perhaps a six-man rotation could have helped alleviate the stress, not only on Tanaka, but all the other pitchers in the rotation, and all across baseball.
……
Jimenez, Corey Kluber, Trevor Bauer and T.J. House. The Cleveland Indians have been at the top of the AL in ERA since May 25, a stretch of 29 straight weeks.
Over all, he is 38-23 with a 3.24 E.R.A., averaging 11.2 strikeouts per nine innings. Last year, he led the major leagues with 277 strikeouts.
config_dict: {
    sentence_labels: [0, 0, 0, 0, 1, 1, 1, 0, ……, 1, 1, 1, 1, 0, 0],
    number_of_segments: 3,
    do_top_p: false,
    do_top_k: true,
    top_k: 40,
    model_name: EleutherAI/gpt-j-6B }

Figure 5: **Comparison to SeqXGPT.** (A): A sample from SeqXGPT (Wang et al., 2023a); (B): A sample generated for our MGT localization task. We observe that SeqXGPT primarily focuses on scenarios where the first part of an article is human-authored, and all subsequent sections are generated by LLMs. In contrast, our approach handles a more realistic and complex scenario, where multiple segments in an article are generated by LLMs. Additionally, our synthetic data provides more fine-grained annotations, including sentence-level labels, LLM sampling strategy, number of machine-generated segments, among others. See Appendix B.2 for discussion.

Figure 6: **An article analyzed by GPT-Zero.** GPT-Zero inaccurately identified the manipulated article as human-generated and consequently label all sentences as human-written. See Appendix B.3 for details.

Model Scale	GPT-2 -1.5B	GPT-Neo -2.7B	OPT -2.7B	GPT-J -6B	GPT-NeoX -20B	mAP	All
Roberta-L	57.24	58.05	49.38	47.41	41.32	50.67	50.85
Roberta-L+vote	71.03	72.03	64.37	62.28	55.39	65.02	65.50
Roberta-L+AdaLoc(skip)	72.07	72.01	67.64	65.79	60.43	67.59	67.63
Roberta-L+AdaLoc(middle)	74.92	74.84	69.85	68.87	64.21	70.54	70.63
Roberta-L+AdaLoc	**82.82**	**82.46**	**78.69**	**76.90**	**71.62**	**78.49**	**79.13**

Table 8: **Ablation Study on GoodNews (Biten et al., 2019).** "skip" denotes that there is no overlapping between different chunks. "middle" means we leverage AdaLoc to predict whether the sentence in the middle is machine-generated. By default, AdaLoc is combined with our majority vote strategy. See Appendix B.4 for discussion.