Spotting AI's Touch: Identifying LLM-Paraphrased Spans in Text

Anonymous ACL submission

Abstract

AI-generated text detection has attracted increasing attention as powerful language models approach human-level generation. Limited work is devoted to detecting (partially) AI-paraphrased texts. However, AI paraphrasing is commonly employed in various application scenarios for text refinement and diversity. To this end, we propose a novel detection framework, paraphrased text span detection (PTD), aiming to identify paraphrased text spans within a text. Different from text-level detection, PTD takes in the full text and assigns each of the sentences with a score indicating the paraphrasing degree. We construct a dedicated dataset, **PASTED**, for **pa**raphra**s**ed **te**xt span **d**etection. Both in-distribution and out-of-distribution results demonstrate the effectiveness of PTD models in identifying AI-paraphrased text spans. Statistical and model analysis explains the crucial role of the surrounding context of the paraphrased text spans. Extensive experiments show that PTD models can generalize to versatile paraphrasing prompts and multiple paraphrased text spans.

1 Introduction

Recent advances in large language models (LLMs) (Touvron et al., 2023; Brown et al., 2020; OpenAI, 2023b) have raised concerns about potential misuse, including student plagiarism and the spread of fake news (Mitchell et al., 2023). A line of work (OpenAI, 2023a; Li et al., 2023b; Mitchell et al., 2023; Yang et al., 2023b) focuses on AI-generated text detection, which assigns a label of "human-written" or "machine-generated" to a text. In addition to pristine AI-generated texts, AI paraphrasing is frequently utilized to polish writings or enhance textual diversity. However, there is limited research on fine-grained detection of texts partially paraphrased or polished by AI. Despite utilizing human-written texts as a base, AI paraphrasing can also suffer from AI-generation issues such as hallucination (Ji et al., 2023) and bias (Navigli et al., 2023), posing significant challenges to achieving trustworthy AI. For example, strict censorship of AI paraphrasing should be implemented in education to ensure factual and harmless content.

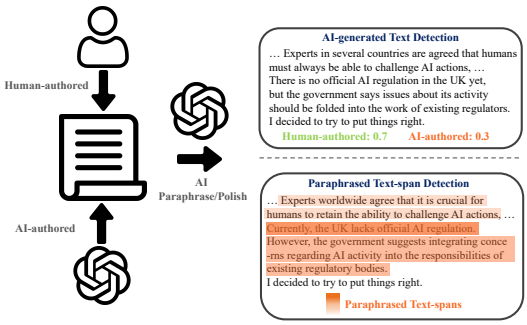

Figure 1: A comparison between AI-generated text detection and paraphrased text span detection, which identifies paraphrased text spans with paraphrasing degree informed, i.e., darker colors denote larger differences between the original and paraphrased text spans.

To this end, we propose a new task called paraphrased text span detection (PTD). PTD predicts a score for each sentence within a long text, identifying AI paraphrased spans comprising consecutive sentences. The sentence-level score also reflects the degree to which the paraphrased text deviates from the original text, i.e., *paraphrasing degree*, as shown in Figure 1. Since there is no existing data for training such a detection system, we assemble a dataset called **PASTED** (**pa**raphrased text **s**pan de**te**ction **d**ataset), including original texts and corresponding paraphrased texts. We obtain these paraphrases by paraphrasing parts of the original texts while keeping other sections intact. The original texts encompass not only human-written compositions but also machine-generated texts. Our goal is for a model trained on our dataset to detect AI-paraphrased texts from new domains and unseen models. To achieve this, we construct an

additional generalization testset to evaluate out-of-distribution (OOD) performance, where texts are paraphrased by a novel paraphraser with different prompts.

The PTD approach is based on the observation that AI-paraphrased text exhibits distinct writing patterns compared to both the original text and its surrounding context. This observation is supported by statistical findings and model analysis. Formally, a PTD model encodes the full text and assigns a sequence of labels (classification model) or scores (regression model) to all sentences in the text. For classification, the labels indicate whether a sentence is paraphrased or not. For regression, the scores quantify the degree of paraphrasing by calculating the difference between each paraphrased sentence and its aligned original text span. To construct reference scores, a dedicated algorithm is devised for aligning paraphrasing pairs.

In-distribution performance shows that all methods effectively distinguish paraphrased text spans, with AUROC exceeding 0.95. While the classification model achieves higher detection accuracy, regression models provide more accurate predictions of the degree of paraphrasing, aligning with OOD performance. The aggregate regression model, which considers all types of divergences, achieves the best overall performance. Identifying context-aware paraphrases poses greater difficulty. On the other hand, models obtain better performance on texts with more paraphrased sentences. Results from the generalization testset demonstrate that all methods can generalize to OOD texts and various paraphrasing prompts despite a performance decrease compared with ID.

Analysis reveals that while paraphrased text displays distinct writing patterns, the surrounding context in the original text significantly plays a crucial role in model detection. Empirical results demonstrate that all models can generalize to texts with multiple paraphrased spans, despite being trained on texts with only one paraphrased span. Moreover, our PTD models can robustly resist minor-perturbed texts that should not be regarded as paraphrases, resulting in few misclassifications. Lastly, we demonstrate the effectiveness of our PTD model in defending against paraphrasing attacks and protecting traditional AI-generation detection systems. The data and code are available at `https://anonymous.com`.

2 Related Work

AI-generated text (AI-generation) detection has received increasing attention. Li et al. (2023b); Chakraborty et al. (2023) systematically discuss the differentiability of AI texts. Various features are explored for detetcion, including n-gram features (Badaskar et al., 2008), entropy (Lavergne et al., 2008; Gehrmann et al., 2019), perplexity (Beresneva, 2016) and model-wise features (Li et al., 2023a). A more direct method involves training neural classifers (Bakhtin et al., 2019; Fagni et al., 2021; Uchendu et al., 2020; OpenAI, 2023a; Li et al., 2023b). Adversarial learning is utilized for robustly detecting AI generations (Hu et al., 2023; Koike et al., 2023b). Another approach proposes training-free methods for detecting AI generation. Methods by Mitchell et al. (2023); Bao et al. (2023) utilize negative curvature regions in the log probability of a model to identify machine-generated text. Additionally, Yang et al. (2023b) compare n-gram features between human-written and AI-generated continuations of text. Different from text-level detection, we propose a more fine-grained approach that identifies paraphrased sentences within a larger body of text. Sentence-level AI-generation detection proposed by Wang et al. (2023) constructs data by breaking long generations into sentences and trains sequence labeling models accordingly. In contrast, we propose detecting paraphrased spans, which can consist of one or multiple sentences within longer texts, focusing on the distinct paraphrasing patterns compared with the original text and the surrounding context. Sadasivan et al. (2023) propose that AI-generation detection can be vulnerable against paraphrasing attacks. Tripto et al. (2023) discuss the authorship of AI-paraphrased texts. Krishna et al. (2023) utilize retrieval to assist detectors to defend against paraphrasing attacks, while Yang et al. (2023a) construct a dedicated dataset with ChatGPT polished texts to increase detection robustness. In contrast, we present a novel framework that identifies AI-paraphrased text spans within a given text and reflects the degree of paraphrasing for each sentence.

3 Problem Definition

3.1 AI-generated Text Detection

Given a text sequence x, the AI-generated text detection model predicts a label y of "human-written" or "machine-generated" based on a probability distribution: $p(y|x)$.

3.2 Paraphrased Text Span Detection

A piece of text x can be segmented into a series of sentences: $\{s_1, s_2, \cdots, s_n\}$. A Paraphrased Text Span Detection (PTD) model is optimized to predict a sequence of binary labels $\{c_1, c_2, \cdots, c_n\}$ or continuous scores $\{r_1, r_2, \cdots, r_n\}$ for these sentences, identifying whether each sentence s_i has been paraphrased by an AI model given the full text x. For each sentence s_i within the text x, the PTD model is tasked with either: (1) Producing a probability distribution $p(c_i|x, i)$ over binary labels, where c_i indicates whether the sentence s_i is paraphrased; (2) Outputing a continuous score $e(x, i)$ that represents the extent of paraphrasing for the sentence s_i. In addition to detection, the model is expected to provide insights into the degree of deviation of the paraphrased text from the original text, quantifying changes in meaning, structure, or other linguistic features that signify paraphrasing.

4 Dataset Construction

In general, PASTED consists of in-distribution training, validation and test sets, along with a generalization testset. We first randomly collect 10% of the original texts from the AI-generation detection dataset (Li et al., 2023b) to collect original texts which encompasses various writing tasks and large language models. The original texts can be either authored by humans or AI, both of which have practical applications. Paraphrasing human-authored compositions can infringe upon composition copyright, while effective paraphrasing can help machine-generated news evade detection. We consider two paraphrasing styles: *context-agnostic paraphrasing* and *context-aware paraphrasing*. Context-agnostic paraphrasing modifies texts without considering the surrounding context and is more commonly used. Context-aware paraphrasing considers the context, bringing larger challenges to detection, as the paraphrases are more coherent and consistent with the context.

In-distribution Data. To simulate real-world scenarios, we employ a sampling process that randomly paraphrases a text-span of several consecutive sentences in the original text. The selected text span consists of 1 to 10 sentences. For context-agnostic paraphrasing, we use a powerful commercial LLM (ChatGPT (Brown et al., 2020)) to construct paraphrases given the independent candidate text span without considering the context. For context-aware paraphrasing, we consider Dipper (Krishna et al., 2023), an 11B model which supports paraphrasing conditioned on the surrounding context. We present several data cases in Appendix C. We conduct paraphrasing on both human and machine texts and collect 83,089 instances (28,473 original texts and 54,616 paraphrased texts) after text pre-processing and filtering. We split the data into train/validation/test sets, with an 80%/10%/10% partition. Detailed data statistics can be referred to in Appendix B.

Generalization Testset. In addition to the in-distribution testset, we construct an additional generalization testset where texts are paraphrased by a novel LLM with different paraphrasing prompts. We employ the same sampling process to generate paraphrases on the out-of-distribution testset from Li et al. (2023b). Recent research (Koike et al., 2023a; Kumarage et al., 2023; Lu et al., 2023) demonstrates that prompt engineering can be employed to evade detection effectively. To this end, we utilize an unseen paraphraser, GPT-4 (OpenAI, 2023b), and explore various prompt variants with increasingly complex instructions for generating elaborate paraphrases. The prompts used for data construction are presented in Appendix D. Ultimately, our OOD evaluation comprises a total of 9,372 instances (1,562 original texts and 7,810 paraphrased texts).

5 Method

A PTD model first decomposes the given text into sentences and predicts a label or score for each sentence based on the full text.

5.1 Sentence-level Classification

We treat each paraphrased sentence in the paraphrased text span as "paraphrased" and others as "original", utilizing cross-entropy (CE) to optimize the model.

$$\mathcal{L}_{\text{CE}}(\theta) = -\sum_{i=1}^{n}[c_i \log(p_\theta(c_i|x, i)) + (1 - c_i) \log(1 - p_\theta(c_i|x, i))] \quad (1)$$

where θ denotes the model parameters.

Limitations of Classification. A major issue of classification arises from its assumption that all paraphrased sentences are equally labeled as "paraphrased", which can overlook the varying degrees

of difference involved in paraphrasing. For instance, when given the sentence "I lost my keys yesterday", the model receives the same label for both paraphrases "Yesterday I lost my keys" and "Yesterday my keys went missing". However, the latter exhibits disparities with the original sentence in terms of word choices and syntax structure. This lack of calibration in predicted probabilities can reduce reliability and interpretability. Therefore, classification models are less resistant to minor text perturbations which should not be regarded as paraphrases (Section 9).

5.2 Sentence-level Regression

To this end, we propose span-level regression, which leverages various difference quantification metrics to inform to what extent each text span has been modified during paraphrasing. Instead of assigning labels to sentences, our model is trained to predict a difference score r_i. This score is calculated using a difference scoring function $f(s_i, t_i)$, such as BLEU score (Papineni et al., 2002). Here, t_i represents the text span in the original text that aligns with the current paraphrased sentence s_i. For training on the original text in our dataset, t_i simply refers to s_i itself. By computing difference scores for each sentence, we optimize our model using mean squared error loss:

$$\mathcal{L}_{\text{MSE}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(f(s_i, t_i) - e_\theta(x, i))^2 \quad (2)$$

Aligning Paraphrased Text Spans. Accurately aligning paraphrased sentences with the original text spans is necessary for reliable indication of the paraphrasing degree as regression labels. A major challenge arises when paraphrasing a text span containing consecutive sentences, which can result in a different number of sentences (41%/45%/14% ratio for fewer, equal, and more sentences in paraphrased text). Furthermore, some paraphrased texts involve reordering at the sentence level. We provide a case illustration in Appendix A. To this end, we propose an alignment algorithm in light of sentence similarity (Reimers and Gurevych, 2019) to align paraphrased sentences with their original counterparts. Our approach involves greedily traversing each paraphrased sentence and identifying a span of consecutive original sentences that share a high semantic similarity. If no suitable span is found, we resort to finding the most semantically similar original sentence (Appendix A).

Difference Quantification Function. We consider a range of different functions to quantify the paraphrasing degree. Given the aligned text span pairs, the most straightforward metric is the *lexical divergence* between them, which can be measured using common similarity-based metrics, e.g., BLEU (Papineni et al., 2002). To capture *grammatical divergence*, we also consider the text similarity score of the part-of-speech sequences. We subtract the BLEU score from 1 to denote the divergence score. For a more comprehensive measure of *syntactic divergence*, we can calculate the tree edit distance between syntax trees (Zhang and Shasha, 1989) at the third level (Bandel et al., 2022). To normalize the tree edit distance, we divide it by the maximum number of nodes in both trees. This normalization results in a difference score ranging from 0 to 1, where 0 indicates identical trees. Finally, we can aggregate all these divergence metrics, which measure the paraphrasing degree from different granularities and views. Specifically, we train the regression model to fit a *aggregate divergence* function encompassing a set of metrics $\{f_1, f_2, \ldots, f_d\}$:

$$\hat{\mathcal{L}}_{\text{MSE}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}(f_j(s_i, t_i) - e_\theta(x, i)_j)^2 \quad (3)$$

The scores of each dimension are averaged as the final aggregate score.

6 Experiment Setup

Settings. We tokenize text into sentences using the NLTK sentence tokenizer (Bird et al., 2009). To perform context-agnostic paraphrasing, we utilize the GPT-3.5-Trubo API. For context-aware paraphrasing with Dipper, we use the default setting with lexical diversity set at 80 and order diversity set at 60. For the generalization testset, we use GPT4 API "gpt-4-1106-preview". To measure sentence similarity in aligning paraphrased sentences, we employ a sentence-transformers model[1] (Reimers and Gurevych, 2019). For constructing regression labels, we obtain part-of-speech tags and constituency parses using the Stanza parser (Qi et al., 2020). We quantify lexical and POS divergence using a 4-gram sentence-level BLEU score and calculate tree edit distance with the ZSS algorithm (Zhang and Shasha, 1989). Fol-

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

lowing Li et al. (2023b), we train classifiers and regression models based on Longformer (Beltagy et al., 2020) by adding a linear layer. All models are trained for 2 epochs on 1 V100 GPU with a batch size of 12. We used the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.005 and set the dropout rate at 0.1. We use GPT-2 large (Radford et al., 2019) to compute text perplexity for experiments in Secition 7 and Section 9.

Evaluation Metrics. To assess the performance of paraphrased sentence detection, we utilize two metrics: **AUROC** (Area Under the Receiver Operating Characteristic curve) and detection accuracy. AUROC measures a classifier's capability to differentiate between positive and negative classes, with a value of 1.0 indicating perfect classification and 0.5 representing random guessing. Following Krishna et al. (2023), we fix the false positive rate (FPR) of 1% and adjust the decision boundary accordingly to report accuracy, ensuring that human-authored text is rarely classified as machine-generated. We denote this metric as **Accuracy (FPR 1%)**. To assess the estimation accuracy of paraphrasing degree, we calculate the Pearson Correlation between model-predicted scores and reference scores. For reference scores, we adopt the lexical diversity and syntactic diversity proposed by Bandel et al. (2022), quantifying lexical and syntactic diversity in generated paraphrases. A high correlation indicates that the model accurately predicts differences between paraphrases and the corresponding original sentences, i.e., the paraphrasing degree. We denote these two correlation scores as **Lexical Corr.** and **Syntactic Corr.**, respectively. For classification models, we utilize the prediction confidence of the classification model as a proxy for predicting paraphrasing degree.

7 Understanding Paraphrased Compositions

We statistically compare original texts and paraphrased texts to assess their distinguishability. Initially, we examine whether paraphrases display distinct word distributions. We randomly divide the dataset into two halves and calculate the Kullback-Leibler (KL) divergence of the top 100 word frequency distribution between original texts and paraphrased texts. We average results across five seeds to reduce randomness. The KL divergence between the original texts and partially paraphrased texts (0.0018) is significantly lower than that between the paraphrased text spans and the corresponding original text spans (0.056). In other words, when considering the full context, the distinct writing pattern of paraphrasing can be overshadowed, emphasizing the necessity for fine-grained detection.

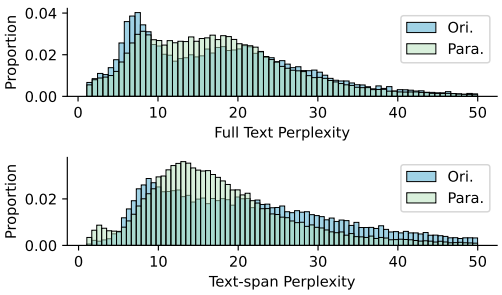

Figure 2: Perplexity distribution of the complete texts and the exact text spans (original v.s. paraphrase).

This finding is also evident in Figure 2, which shows the perplexity distribution. The perplexity distribution of paraphrased text closely aligns with that of original text. Note that the original text can either be sourced from human or machine, which forms isolated perplexity distributions (Mitchell et al., 2023; Li et al., 2023b). Nevertheless, the perplexity of paraphrased text spans exhibits a distribution centered around a value of 12, indicating a strong writing pattern regardless of the data source. Therefore, sentence-level detection captures paraphrasing patterns more precisely compared to text-level detection.

We further compare the original and the paraphrased text spans under different data sources and paraphrasing methods. Results show that both human-written and machine-generated sources yield similar word distribution divergences (0.083 v.s. 0.052). However, context-agnostic paraphrasing results in a significantly different word distribution than context-aware paraphrasing (**0.22** v.s. 0.041), indicating that context-agnostic approaches are more lexically diverse but may largely deviate from the original style. Word clouds for both paraphrasing methods are shown in Appendix E, demonstrating the novel word distribution introduced by context-agnostic paraphrasing.

8 Results

8.1 In-distribution Performance

The detection performance on random-split testsets is presented in the upper part of Table 1. We consider two baselines: (1) Random which calculates BLEU scores of the paraphrased sentence

Model	AUROC↑	Accuracy (FPR 1%)↑	Lexical Corr.↑	Syntactic Corr.↑
Random	0.50	0.00%	0.07	0.07
Oracle	1.00	100.00%	0.88	0.88
In-distribution Detection				
Classification	**0.97**	**69.27%**	0.64	0.67
Regression (lexical)	**0.97**	64.04%	0.69	0.71
Regression (grammatical)	0.96	54.80%	**0.70**	**0.72**
Regression (syntactic)	0.96	47.45%	0.67	**0.72**
Regression (aggregate)	**0.97**	59.45%	**0.70**	**0.72**
Out-of-distribution Detection				
Classification	**0.94**	**47.21%**	0.62	0.66
Regression (lexical)	**0.94**	42.57%	**0.66**	**0.70**
Regression (grammatical)	0.93	20.29%	**0.66**	0.69
Regression (syntactic)	0.90	9.63%	0.60	0.65
Regression (aggregate)	**0.94**	26.21%	**0.66**	**0.70**

Table 1: In-distribution (upper part) and out-of-distribution (lower part) detection performance of classification and regression methods. Accuracy (FPR 1%) refers to the accuracy with a false positive rate maintained under 1%. The lexical and syntactic correlation (Corr.) indicates the accuracy in predicting paraphrasing degree.

with a random sentence in the original text and (2) Oracle which calculates BLEU scores of the paraphrased sentence with the aligned text span. All detection methods effectively distinguish paraphrased text spans, with AUROC exceeding 0.95. Although the classification model performs better in detection, it falls short compared to regression models in predicting paraphrasing degree due to improper calibration (discussed in Section 5.1). In contrast, regression models demonstrate stronger alignment with the reference differences between original texts and paraphrases, both lexically and syntactically. In other words, regression models are more reliable indicators of the extent of difference present in a paraphrased text span. Regression models with grammatical or syntax supervision obtain the best lexical and syntactic correlation. The classification model and lexical regression model obtain the best accuracy (FPR 1%), effectively identifying paraphrases while maintaining a low false positive rate (1%). Overall, the aggregate regression achieves the best in-distribution performance.

Effects of Data Source and Paraphrasing Method. Consistent with the statistical results in Section 7, the data source (human-written or machine-generated) has minimal effect on detection performance, as indicated by an AUROC score of 0.97 for both. In contrast, detecting context-aware paraphrases proves to be more challenging, achieving an AUROC score of 0.94 compared to 0.98 for context-agnostic ones. The detection performance (lexical regression) on text from different domains and LLMs is presented in Figure 3. Paraphrased spans in technological news (TLDR) or scientific writings (Sci_Gen) are comparatively challenging to identify, followed by Wikipedia articles (SQuAD). On the other hand, paraphrasing texts produced by encoder-decoder LLMs (FLAN-T5) or larger LLMs (OpenAI, LLaMA, and GLM) pose significantly greater difficulties.

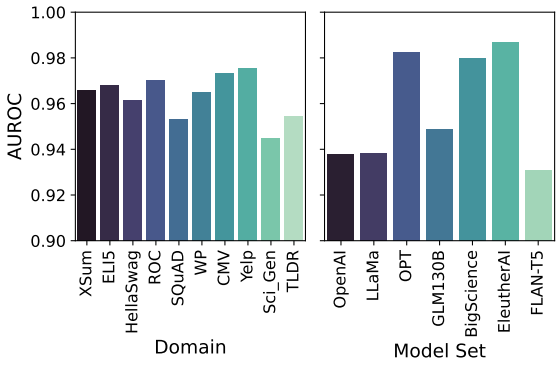

Figure 3: Detection performance (lexical regression) on text from different domains and LLMs.

Effect of Number of Paraphrases. The impact of the number of paraphrased sentences in a text is illustrated in Figure 4. As depicted, the detection difficulty decreases as the text span includes more paraphrased sentences, as indicated by both AUROC and accuracy. This could be attributed to the fact that paraphrasing a longer text span (i.e., with more sentences) can showcase more pronounced paraphrasing styles, encompassing both lexical usage and syntax structure. Consequently, it exhibits a more discernible contrast with its surrounding

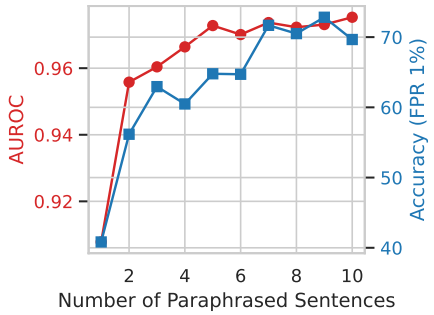

Figure 4: Detection performance (lexical regression) w.r.t. the number of paraphrased sentences in a text.

context. In contrast, when only one sentence is paraphrased, especially if it is short, there are limited features available for detection.

8.2 Out-of-distribution Performance

The lower part of Table 1 presents the out-of-distribution performance on the generalization testset, which consists of paraphrased texts generated by novel LLMs. Despite a degradation in performance across all methods and metrics, they still achieve reasonably good results in terms of both paraphrase identification (ARUOC) and paraphrasing degree prediction (correlation). Similar to the in-distribution results, the classification model demonstrates the highest detection performance with an accuracy of 47.21%. Differently, the classification and lexical regression model are substantially stronger than all other models for detecting paraphrases, with a much higher accuracy. In predicting the degree of paraphrasing, the regression models perform better, particularly the aggregate regression model which achieves lexical and syntactic correlations of 0.66 and 0.70 respectively. We construct the generalization testset using various paraphrasing prompts (Appendix D). Empirical results demonstrate that the detector is resistant to prompt variance, with an AUROC exceeding 0.92 across all prompts. The performance of lexical regression against prompt variance is shown in Appendix G.

9 Analysis

In this section, we set the decision boundary to maintain an **FPR of 1%** on the **validation** set to accommodate various testsets across all analytic experiments and report detection accuracy.

Effect of Surrounding Context. As discussed in Section 7, paraphrased texts exhibit strong writing patterns different from the surrounding texts. We conduct an ablation study on the effect of the surrounding context, with results shown in Table 2. The models in the table are trained merely on the paraphrased sentences, without considering the surrounding context. As shown, both classification and regression (lexical) models suffer substantial performance degradation when missing the context information, with over 100% performance drop on detection accuracy. To gain further insight, we analyze perplexity variations across token positions around the boundary of the paraphrased text span. The results are presented in Figure 5, where we observe perplexity impulses at both boundaries of the paraphrased text span. Context-agnostic paraphrases typically exhibit higher token perplexity from the beginning of the paraphrase. While context-aware paraphrases display lower token perplexity before reaching the end of the paraphrase, they encounter a substantial increase in perplexity afterwards. This indicates that paraphrased sentences cannot perfectly integrate into the original text, even if context is considered during paraphrasing.

Model	AUROC	Accuracy (FPR 1%)
Classification	0.87(-0.10)	33.39%(-35.88%)
R-lexical	0.86(-0.11)	31.01%(-33.03%)

Table 2: Detection performance without considering the surrounding context. "R" stands for "Regression".

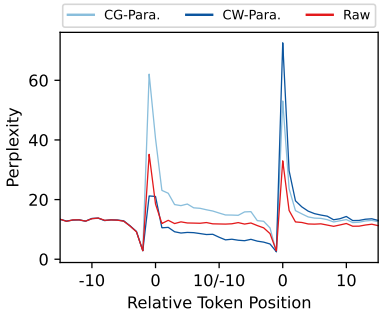

Figure 5: Token perplexity around the boundary of the paraphrased texts: two "0" denote the start and the end of the paraphrased text span. The x-axis represents the relative token position with respect to each boundary.

Generalization to Multiple Paraphrased Text Spans. Our training data only considers paraphrasing one text span in a text. In many application scenarios, users can paraphrase multiple text spans. To this end, we construct an additional testset where multiple text spans are paraphrased

Model	AUROC	Accuracy
Classification	**0.96**	**67.68**
R-lexical	0.93	66.30
R-grammtical	0.93	62.75
R-syntactic	0.93	53.11
R-aggregate	0.94	64.76

Table 3: Detection performance of generalization to texts with multiple paraphrased text-spans.

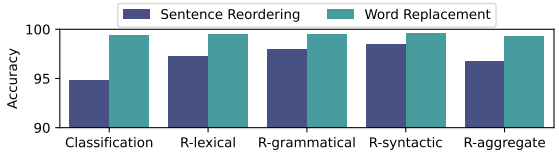

Figure 6: Detection robustness towards attacks by constructing misleading texts with minor modifications.

within each text. The testset consists of 500 randomly sampled in-distribution test instances. For each text, we randomly choose 2 to 5 non-adjacent text spans which consist of 1 to 3 sentences, and paraphrase these text spans using ChatGPT. The results are shown in Table 3. Although all detection methods experience a performance decline in terms of AUROC, they achieve a comparable detection accuracy, demonstrating the generalization ability to text with multiple paraphrased text spans. We present a case study of PTD in Appendix F.

Robustness to Minor Text Modification. A desirable characteristic of a paraphrase detector, in addition to detection accuracy, is the ability to distinguish texts with minor modifications from true paraphrases. We consider two types of minor modifications: sentence reordering and word replacement. For sentence reordering, we utilize Dipper with maximum ordering control and minimum lexical control. For text replacement, we randomly mask 10% of the text and use T5-3B (Raffel et al., 2020) to fill in the masked blank. Data details can be referred to in Appendix H. We evaluate these modifications on the in-distribution testset by considering all minor-perturbed texts as "non-paraphrased". As shown in Figure 6, all methods rarely misidentify texts with word replacements as paraphrased, with nearly perfect prediction accuracy. In contrast, texts with sentential reordering pose greater confusion and yield considerably lower accuracy. Notably, regression models incorporating grammatical or syntax information exhibit more resistance to such confusions compared to the classification model which performs worst.

Model	HumanRec	MachineRec	AvgRec
Detector	88.78%	37.05%	62.92%
+ Defender	88.98 %	78.50%	83.74%

Table 4: AI-generated detection performance against paraphrasing attacks with paraphrased text-span detection as a defense.

Defending AI-generation Detection. Previous work (Krishna et al., 2023; Sadasivan et al., 2023) has shown that paraphrasing attacks significantly degrade AI-generation detection systems. We propose a two-stage detection method, which utilizes paraphrased text span detectors as a pre-defense mechanism to block out paraphrased texts and employs traditional AI-generation detectors if no paraphrasing is detected. We calculate the text score by averaging the predicted scores of all sentences and evaluate the model on the paraphrasing attack testset proposed by Li et al. (2023b). We use an off-the-shelf AI-generation detector[2] and employ the aggregate regression model for defense. The results are presented in Table 4, where it can be observed that most machine-generated texts were misclassified by the AI-generation detector due to paraphrasing, resulting in low MachineRec (recall on machine-generated texts). The paraphrasing indicator successfully identifies most of the paraphrased texts and significantly improves averaged recall (AvgRec) scores while maintaining high recall scores on human texts (HumanRec).

10 Conclusion

In this work, we propose a detection framework, paraphrased text span detection (PTD), which aims to identify text spans paraphrased by AI, from a long text. We built a dedicated dataset, PASTED, based on which we train classification and regression models for PTD. Both in-distribution and out-of-distribution results demonstrate that our methods can effectively detect paraphrased text spans. Although classification models achieve better detection accuracy, they fall behind regression models in predicting the paraphrasing degree. Statistical and model analysis showcases the importance of the context surrounding the paraphrased text spans for detection performance. Extensive experiments demonstrate the generalization to paraphrasing prompt types and multiple paraphrased text spans.

[2]https://github.com/yafuly/DeepfakeTextDetect

Limitations

Although we extensively experiment and analyze the implementation of our newly proposed task PTD using the newly established dataset PASTED, there are several limitations: (1) We consider both context-agnostic and context-aware paraphrasing using limited paraphrasers and prompts. Future work should focus on constructing more challenging paraphrases. (2) We implement effective detection methods based on Longformer, but more advanced backbones like LLaMA can be explored. (3) To simulate real-life applications, we randomly paraphrase text spans in existing datasets. Future work should aim to construct more realistic data by involving crowdsourcing.

Ethical Considerations

We honor the Code of Ethics. No private data or non-public information is used in this work. We adhere to the terms of companies offering commercial LLM APIs and express our gratitude to all global collaborators for their assistance in utilizing these APIs.

References

Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. 2008. Identifying real or fake articles: Towards better language modeling. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. 2022. Quality controlled paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 596–609. Association for Computational Linguistics.

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.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Daria Beresneva. 2016. Computer-generated text detection using machine learning: A systematic review. In *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings 21*, pages 421–426. Springer.

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

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, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey 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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of ai-generated text detection. *CoRR*, abs/2304.04736.

T Fagni, F Falchi, M Gambini, A Martella, M Tesconi, et al. 2021. Tweepfake: About detecting deepfake tweets. *PLOS ONE*, 16(5):1–16.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. GLTR: statistical detection and visualization of generated text. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 111–116. Association for Computational Linguistics.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. RADAR: robust ai-text detection via adversarial learning. *CoRR*, abs/2307.03838.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2023a. How you prompt matters! even task-oriented constraints in instructions affect llm-generated text detection. *CoRR*, abs/2311.08369.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2023b. OUTFOX: llm-generated essay detection through in-context learning with adversarially generated examples. *CoRR*, abs/2307.11729.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *CoRR*, abs/2303.13408.

Tharindu Kumarage, Paras Sheth, Raha Moraffah, Joshua Garland, and Huan Liu. 2023. How reliable are ai-generated-text detectors? an assessment framework using evasive soft prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1337–1349. Association for Computational Linguistics.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. 2008. Detecting fake content with relative entropy scoring. *PAN*, 8:27–31.

Linyang Li, Pengyu Wang, Ke Ren, Tianxiang Sun, and Xipeng Qiu. 2023a. Origin tracing and detecting of llms. *CoRR*, abs/2304.14072.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023b. Deepfake text detection in the wild. *CoRR*, abs/2305.13242.

Ning Lu, Shengcai Liu, Rui He, Qi Wang, and Ke Tang. 2023. Large language models can be guided to evade ai-generated text detection. *CoRR*, abs/2305.10847.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *CoRR*, abs/2301.11305.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *ACM J. Data Inf. Qual.*, 15(2):10:1–10:21.

OpenAI. 2023a. Ai text classifier.

OpenAI. 2023b. GPT-4 technical report. *CoRR*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 101–108. Association for Computational Linguistics.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

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.

Nafis Irtiza Tripto, Saranya Venkatraman, Dominik Macko, Róbert Móro, Ivan Srba, Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. A ship of theseus: Curious cases of paraphrasing in llm-generated texts. *CoRR*, abs/2311.08374.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. Seqxgpt: Sentence-level ai-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1144–1156. Association for Computational Linguistics.

Lingyi Yang, Feng Jiang, and Haizhou Li. 2023a. Is chatgpt involved in texts? measure the polish ratio to detect chatgpt-generated text. *CoRR*, abs/2307.11380.

Xianjun Yang, Wei Cheng, Linda R. Petzold, William Yang Wang, and Haifeng Chen. 2023b. DNA-GPT: divergent n-gram analysis for training-free detection of gpt-generated text. *CoRR*, abs/2305.17359.

Kaizhong Zhang and Dennis E. Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, 18(6):1245–1262.

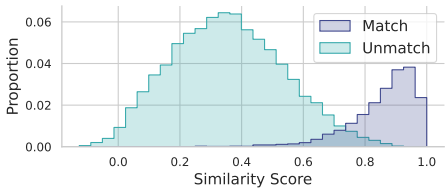

Figure 7: Similarity score distribution of paraphrases and non-paraphrases.

A Aligning Paraphrased Sentences

We show a case where paraphrasing text involves sentence reordering and many-to-one sentence alignment in Figure 8. The detailed implementation for aligning paraphrased sentences with their reference original sentences is presented in Algorithm 1. Our approach involves greedily traversing each paraphrased sentence and identifying a span of consecutive original sentences that share high semantic similarity. If no suitable span is found, we fallback to finding the most semantically similar original sentence. To determine a similarity threshold, we sample 1,000 instances where the number of sentences remains unchanged during paraphrasing, where most cases exhibit trivial one-on-one alignment upon manual inspection. We compute the semantic similarity between all pairs of paraphrases and original sentences. Pairs with matching indices (e.g, 0 v.s. 0) are considered matched while others are unmatched (e.g, 0 v.s. 1 and 2). The distribution of similarities is depicted in Figure 7, suggesting that 0.75 is a promising threshold value.

B Data Statistics

The data statistics of PASTED dataset is presented in Table 5.

C Data Sample

We present several data samples in Table 6 and Table 7.

D Prompt Design

The prompt templates used for constructing the data set is shown in Figure 10.

E Word Cloud of Paraphrases

The word cloud of word distribution for original texts, context-agnostic paraphrases and context-aware paraphrases is shown in Figure 11.

Algorithm 1 Paraphrase Alignment

Setup: n: number of paraphrased sentences
m: number of reference sentences
$avg(L, i, j)$: calculate mean value of L[i,i+1,...j-1]

Input: mat: Similarity matrix between paraphrased sentences and reference sentences, $R^{n \times m}$
τ: threshold

{initializing}
$A \leftarrow \emptyset$
$i \leftarrow 0$
{align each paraphrased sentence individually.}
while $i < n$ **do**
  {get the similarity of the most similar reference sentence}
  $V_{max} \leftarrow mat[i].max()$
  if $V_{max} \leq \tau$ **then**
    {if the max similarity is less than or equal to the threshold, we just align the paraphrased sentence i with the most similar reference.}
    $idx \leftarrow argmax(mat[i])$
    $A.add((i, (idx, idx + 1)))$
  else
    {If the maximum similarity exceeds the threshold, we align paraphrased sentence i with the longest reference sentence span that has an average similarity greater than the threshold.}
    $W_{size} \leftarrow m$
    $Flag \leftarrow 0$
    while $W_{size} \geq 1$ **do**
      $j \leftarrow 0$
      while $j \leq m - W_{size}$ **do**
        $V_{mean} \leftarrow avg(mat[i], j, j + W_{size})$
        if $V_{mean} > \tau$ **then**
          $A.add((i, (j, j + W_{size})))$
          $Flag \leftarrow 1$
          break
        end if
        $j \leftarrow j + 1$
      end while
      if $Flag = 1$ **then**
        break
      end if
      $W_{size} \leftarrow W_{size} - 1$
    end while
  end if
  $i \leftarrow i + 1$
end while
return A

F Case Study

We present several cases of paraphrased text span detection (aggregate regression model) in Figure 12, Figure 13 and Figure 14.

G Effect of Prompts

The effect of prompts for constructing paraphrases is shown in Figure 9, where "CG" and "CW" are abbreviations for "context-agnostic" and "context-aware" respectively.

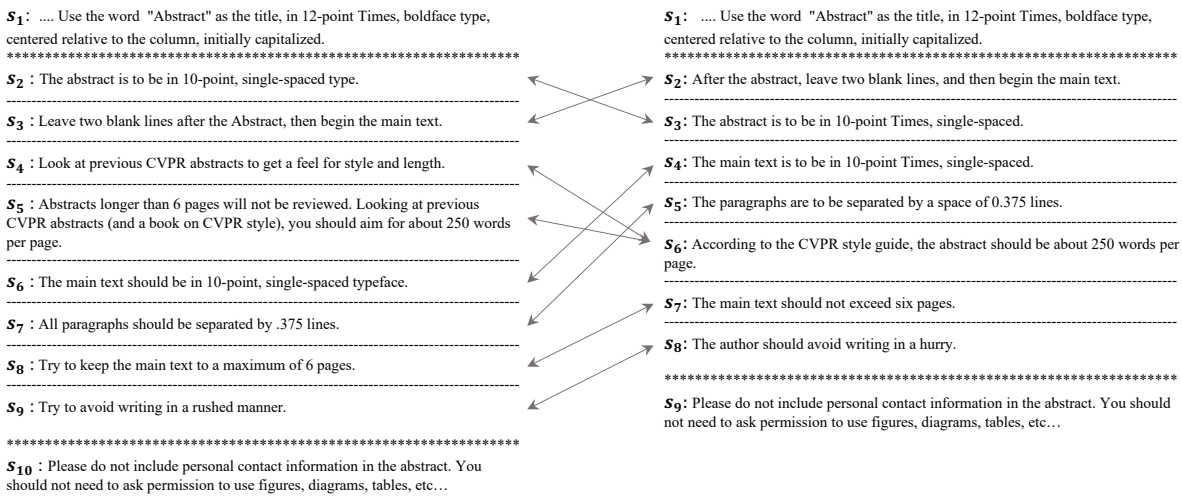

Figure 8: A case of paraphrasing involving sentence reordering and many-to-one sentence alignment is presented. The original text exhibits changes in the order of s_2 and s_3, as well as the order of s_4s_5 and s_6s_7, after paraphrasing. In particular, the text span s_4s_5 in the original text aligns with s_6 in the paraphrased text.

Data Source	Human	Machine	All
Original Texts	6577/857/838	16082/2038/2081	22659/2895/2919
Context-agnostic Paraphrases (ChatGPT)	6577/857/838	16082/2038/2081	22659/2895/2919
Context-aware Paraphrases (Dipper)	6235/830/796	14477/1937/1868	20712/2767/2664
All	20712/2767/2664	46641/6013/6030	66030/8557/8502

Table 5: In-distribution data statistics. The number of train, validation and test set is delimited by "/".

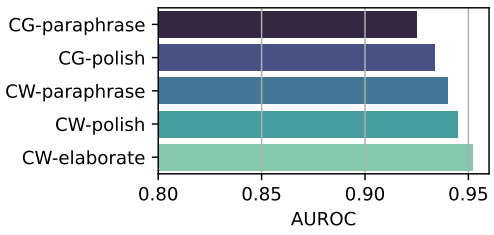

Figure 9: Effect of prompt types on detection performance (lexical regression). "CG" stands for "context-agnostic" and "CW" stands for context-aware. Prompt type instances can be referred to in Appendix D.

H Constructing Texts with Minor Modification

For sentence reordering, we set the lexical control of Dipper to the lowest value (i.e., 20) and the order control value to the max (i.e., 100)[3]. For word replacement, we randomly mask 10% of the tokens and use T5-3B model to fill in these blanks. We apply both methods on the in-distribution testset. To ensure minimal modifications, we calculate a similarity score between the original and paraphrased text. We then establish a threshold to exclude paraphrases that significantly differ from the original texts. We only retain paraphrases with BLEU scores above 70 when compared with their respective original texts. In this way, we construct 273 and 1,310 instances for sentence reordering and word replacement, respectively. We show several cases in Table 8.

[3]Please refer to https://github.com/martiansideofthemoon/ai-detection-paraphrases for details

Raw Text 1	In great anticipation, we made our way from our hotel to the restaurant. We were there at 5:45 and had no problems being seated right away. We had the 14 nam kao tod meat salad, garlic prawns, tom kha gai soup and crispy duck on drunken noodles. It was a lot of food, enough for a doggie bag for later too. To drink, Thai ice tea (creamy almost like an iced coffee) and a Thai Lime drink (refreshing but too sweet, ask for less sugar). Total came to around 75.00 for the two of us. Now, I may be crucified for this but, it wasn't knock my socks off amazing. It was delicious, but I've been spoiled with Thai food here in Toronto. We go to Express Thai on Dufferin which is owned and operated by a Thai family, and I feel the flavours there are just as fresh and authentic as Lotus of Siam.
CG-Para. 1	In great anticipation, we made our way from our hotel to the restaurant. We were there at 5:45 and had no problems being seated right away. We ordered various dishes including the 14 nam kao tod meat salad, garlic prawns, tom kha gai soup, and crispy duck on drunken noodles. The portion sizes were generous, so we even had leftovers for later. For our beverages, we had Thai ice tea, which had a creamy texture similar to iced coffee, and a Thai Lime drink, which was refreshing but slightly too sweet. The bill totaled around $75. While some might strongly disagree, I personally didn't find the food to be exceptionally outstanding. It was delicious, but I've been spoiled with Thai food here in Toronto. We go to Express Thai on Dufferin which is owned and operated by a Thai family, and I feel the flavours there are just as fresh and authentic as Lotus of Siam.
CA-Para. 1	In great anticipation, we made our way from our hotel to the restaurant. We were there at 5:45 and had no problems being seated right away. We ordered the nam kao tod, a meat salad, garlic shrimp, tom yam goong, and duck with drunken noodles. We also had a Thai iced tea and a lime juice, both of which were very sweet. I'm afraid I'm going to be criticized for saying this, but the food wasn't that great. It was good, but not great. The bill came to about $70 for the two of us. It was delicious, but I've been spoiled with Thai food here in Toronto. We go to Express Thai on Dufferin which is owned and operated by a Thai family, and I feel the flavours there are just as fresh and authentic as Lotus of Siam.
Raw Text 2	My friend and I thought it would be fun to try Pub 1842 for a light lunch, considering all we do is eat when we go to Vegas. I can't lie, my friend has an obsession with buffets (MGM buffet is the best) but I was trying to get her to eat something different and on the fancy side. In the middle of the MGM casino, you'll find this restaurant with a great atmosphere. They have a bar area as well, which is perfect for a drink while you're waiting for your table. We sat outside on the patio, which was nice because it was bright and sunny outside. I had the lunch special which consisted of a mixed green salad with candied walnuts, gorgonzola cheese, and dried cranberries. The salad was huge and I was only able to eat half of it. The dressing was on the sweet side, I like mine to be a bit more tangy. The 1842 Burger is the house specialty. I couldn't resist because it had a fried egg on it, and anything with an egg is always a good thing! The sweet potato fries were great! Crispy and perfect on the side. I had the funnel cake for dessert, which was just okay. Not the best funnel cake I've had.
CG-Para. 2	My friend and I thought it would be fun to try Pub 1842 for a light lunch, considering all we do is eat when we go to Vegas. Honestly, my friend is really obsessed with buffets (specifically the MGM buffet), but I wanted to convince her to try something different and more upscale. Right in the middle of the MGM casino, there's this restaurant with a fantastic atmosphere. They even have a bar area, which is perfect for enjoying a drink while waiting for your table. We decided to sit outside on the patio, which was great because it was sunny and bright outdoors. For my meal, I opted for the lunch special, which included a generous mixed green salad topped with candied walnuts, gorgonzola cheese, and dried cranberries. The salad was enormous, and I could only manage to eat half of it. The dressing was on the sweeter side, although I prefer mine to have a tangy kick. The house specialty here is the 1842 Burger. I couldn't resist because it had a fried egg on it, and anything with an egg is always a good thing! The sweet potato fries were great! Crispy and perfect on the side. I had the funnel cake for dessert, which was just okay. Not the best funnel cake I've had.
CA-Para. 2	My friend and I thought it would be fun to try Pub 1842 for a light lunch, considering all we do is eat when we go to Vegas. In the middle of the MGM Grand Casino, this restaurant has a great atmosphere and a bar that is perfect for a drink while you wait for a table. I won't lie, my friend is obsessed with buffets (the MGM Grand buffet is the best), but I wanted to try something different and a little more upscale. I ordered the salad special, which was a mixed green salad with candied walnuts, gorgonzola cheese, and dried cranberries. It was so big I could only eat half of it. We sat outside on the patio, which was nice because it was sunny. The 1842 burger is the specialty. The salad dressing was a little too sweet for my taste. I couldn't resist because it had a fried egg on it, and anything with an egg is always a good thing! The sweet potato fries were great! Crispy and perfect on the side. I had the funnel cake for dessert, which was just okay. Not the best funnel cake I've had.

Table 6: Data samples in PASTED, where "CG-Para." stands for context-agnostic paraphrasing and "CW-Para." denotes context-aware paraphrasing.

Raw Text 3	Make sure to read the manual and allow yourself a few hours to explore all the options. Once you're comfortable with one, train your fingers to type or swipe quicker. This is a must if you're a power user. You'll get the most out of your smartphone by keeping it updated to the latest version of the software. You may also notice an app or two that isn't working properly. Simply uninstall it. Also keep a close eye on your storage space. If you're running out, you may need to delete some pictures, videos and files. Your smartphone is so much more than a phone.
CG-Para. 3	Make sure to read the manual and allow yourself a few hours to explore all the options. Once you're comfortable with one, train your fingers to type or swipe quicker. If you regularly use your smartphone, it is essential to keep it updated with the latest software version for optimal performance. Additionally, you might come across a few malfunctioning apps, in which case, you can easily remove them. If you regularly use your smartphone, it is Also keep a close eye on your storage space. If you're running out, you may need to delete some pictures, videos and files. Your smartphone is so much more than a phone.
CA-Para. 3	Make sure to read the manual and allow yourself a few hours to explore all the options. Once you're comfortable with one, train your fingers to type or swipe quicker. You'll also notice that some of the apps aren't working properly. If you're a power user, you'll want to keep your phone's operating system up to date. If an app isn't working, just uninstall it. Also keep a close eye on your storage space. If you're running out, you may need to delete some pictures, videos and files. Your smartphone is so much more than a phone.
Raw Text 4	Atticus does not want Jem and Scout to be present at Tom Robinson's trial. No seat is available on the main floor, so by invitation of Rev. Sykes, Jem, Scout, and Dill watch from the colored balcony. Atticus establishes that the accusers - Mayella and her father, Bob Ewell, the town drunk - are lying. It also becomes clear that the friendless Mayella made sexual advances toward Tom, and that her father caught her and beat her. Despite significant evidence of Tom's innocence, the jury convicts him. Jem's faith in justice becomes badly shaken, as is Atticus', when the hapless Tom is shot and killed while trying to escape from prison.
CG-Para. 4	Atticus does not want Jem and Scout to be present at Tom Robinson's trial. No seat is available on the main floor, so by invitation of Rev. Sykes, Jem, Scout, and Dill observe the events from the balcony designated for the colored people. Atticus proves that Mayella and her father, Bob Ewell, who is known as the town drunk, are not telling the truth. It also becomes evident that Mayella, who has no friends, made sexual advances towards Tom, and her father discovered it and violently beat her. Although there is substantial evidence to prove Tom's innocence, the jury unfairly finds him guilty. Jem's faith in justice becomes badly shaken, as is Atticus', when the hapless Tom is shot and killed while trying to escape from prison.
CA-Para. 4	Atticus does not want Jem and Scout to be present at Tom Robinson's trial. No seat is available on the main floor, so by invitation of Rev. Sykes, Jem, Scout, and Dill are allowed to sit in the balcony. Atticus proves that the accusers, Mayella and her father Bob Ewell, are lying. It also becomes clear that Mayella, who has no friends, had been trying to seduce Tom, but her father had caught her and beaten her. Tom Robinson is convicted, despite the overwhelming evidence of his innocence. Jem's faith in justice becomes badly shaken, as is Atticus', when the hapless Tom is shot and killed while trying to escape from prison.

Table 7: Data samples in PASTED, where "CG-Para." stands for context-agnostic paraphrasing and "CW-Para." denotes context-aware paraphrasing.

Prompt Type	Train/Dev/Test	Prompt Type	OOD Testsets
Context-agnostic Paraphrase	Paraphrase the following text: [Insert text here]	Context-aware Polish	Please polish and refine the following excerpt from a larger text, considering its original context. Include both the excerpt and the context below for reference: Excerpt: [Insert excerpt here] Context: [Provide surrounding text or context here] please response paraphrased excerpt as following format: Paraphrased Excerpt: ############## [Paraphrased excerpt here] ##############
Prompt Type	**OOD Testsets**		
Context-agnostic Paraphrase	Paraphrase the following text: [Insert text here]		
Context-agnostic Polish	Polish and refine the following text: [Insert text here]	Context-aware Elaborate Paraphrase	Request for Paraphrasing Assistance: Objective: To obtain a paraphrased version of a specified text excerpt, ensuring adherence to the following critical requirements: 1. Fidelity to Original Meaning: The paraphrased text must preserve the complete meaning of the original excerpt without deviation. 2. Contextual Coherence: The paraphrase should maintain coherence and consistency with the overall context of the larger text from which the excerpt is derived. 3. Sophistication in Language Use: The paraphrasing should be executed with a high degree of linguistic proficiency. The final text must be original enough to evade detection by AI-based paraphrasing tools and anti-fake algorithms. Please provide the following information for the paraphrasing task: Excerpt for Paraphrasing: [Insert excerpt here] Contextual Background: [Provide surrounding text or context here] please response paraphrased excerpt as following format: Paraphrased Excerpt: ############## [Paraphrased excerpt here] ##############
Context-aware Paraphrase	Please paraphrase the following excerpt from a larger text, considering its original context. Include both the excerpt and the context below for reference: Excerpt: [Insert excerpt here] Context: [Provide surrounding text or context here] please response paraphrased excerpt as following format: Paraphrased Excerpt: ############## [Paraphrased excerpt here] ##############		

Figure 10: Case illustration for 5 prompt types used in this paper. "Context-agnostic Paraphrase" is used for constructing the in-distribution train and test set, while the other prompts types are used for constructing the generalization testset.

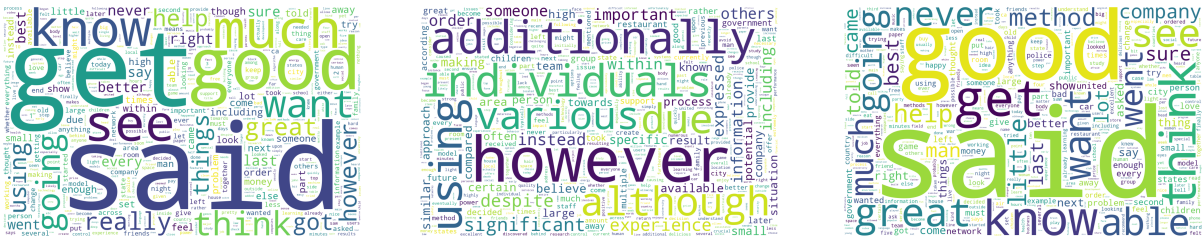

Figure 11: Word clouds of original texts, context-agnostic paraphrases and context-aware paraphrases.

Paraphrased Text

Benjamin Franklin Gates stepped into the chamber and looked around. Where most people saw various symbols of American might and historical jurisprudence, he saw a multitude of hidden signals trying to communicate with him from every corner. "I have come here today in search for information and I do not intend on leaving this room until you tell me everything that my ancestors left behind were lies." He stated before speaking to his shadow. "That would be unwise, Your Highness. Even if they are all lying, there is so much more we could learn about them by studying their words and deeds. Would your Majesty like an example? 'Ours has been A New Day since 1776' was written atop each of our monuments as indicators of America's place among civilized nations. However, it does not appear that any countries within the world have ever seen one such day nor will they see another as long the people continue to live under tyranny!" Benjamin did not pay much attention to the man's argument. Instead, he turned his gaze towards the deteriorating walls adorning the ceiling of the room. "So, what you're saying is that these misleading revelations only serve to divert our attention away from real dangers? What specific danger am I currently confronted with?" Isn't protecting ourselves against Islamic terrorism still important enough to care what someone else says?" He said looking up at his shadow again. The shadow nodded reluctantly "The Crown Prince has always known that those who protect themselves against conflict end up fighting even harder afterwards; thus, training through demonstrations alone serves only to cause unnecessary suffering.My master knows this better then anyone, yet despite this knowledge he has continued to hold onto its teachings with both hands preventing his subjects from learning anything beyond that which is taught in schools throughout his kingdom. If we do not rebel against tyranny, we will either die or face an even worse fate. If you think that the only source of truth is found within school books, then you should never pursue knowledge beyond that, Your Highness. You may think you're being wise but the opposite is really happening. To truly understand who you are, it is essential to comprehend history. It is important to acknowledge that despite challenges, in order to navigate the present, you must learn from past experiences and use them for personal development. History shouldn't solely focus on war and triumphs; hardships and failures also pave the way for progress and the emergence of new ideas. By remembering the courageous fights for freedom and justice, Americans have continuously persevered, regardless of the number of attempts needed to achieve success. They accomplished this while dressed in patriotic red, white, and blue stripes and proudly held flags adorned with stars! Isn't it amazing that all of these occurrences occurred within your lifespan? That's great news, Your Highness.Now that you've had some extra time to reflect on the past let's discuss something less serious shall we?"

Raw Text

Benjamin Franklin Gates stepped into the chamber and looked around. Where most people saw various symbols of American might and historical jurisprudence, he saw a multitude of hidden signals trying to communicate with him from every corner. "I have come here today in search for information and I do not intend on leaving this room until you tell me everything that my ancestors left behind were lies." He stated before speaking to his shadow. "That would be unwise, Your Highness. Even if they are all lying, there is so much more we could learn about them by studying their words and deeds. Would your Majesty like an example? 'Ours has been A New Day since 1776' was written atop each of our monuments as indicators of America's place among civilized nations. However, it does not appear that any countries within the world have ever seen one such day nor will they see another as long the people continue to live under tyranny!" Benjamin didn't give much thought to the man's point but instead looked down at the crumbling walls decorating the chambers ceiling. "You mean these false revelations serve no purpose other than distract us when faced with true threats? What kind of threat am I facing now? Isn't protecting ourselves against Islamic terrorism still important enough to care what someone else says?" He said looking up at his shadow again. The shadow nodded reluctantly "The Crown Prince has always known that those who protect themselves against conflict end up fighting even harder afterwards; thus, training through demonstrations alone serves only to cause unnecessary suffering.My master knows this better then anyone, yet despite this knowledge he has continued to hold onto its teachings with both hands preventing his subjects from learning anything beyond that which is taught in schools throughout his kingdom. We must revolt or perish along side...or perhaps worse in the fight against tyranny. If you believe that truth leads nowhere else to lead then you should never seek out knowledge outside of school books, Your Highness. You may think you're being wise but the opposite is really happening. Only by understanding history can you begin to understand yourself; therefore know this: Come what may, in order to survive your time requires you to embrace the lessons learned as part of personal growth rather than avoiding facts because they make things difficult right now. This is also why history doesn't need to concern itself solely with warfare and victories - struggles and defeats often allow progress to take root making possible new ideas to flourish once gone forgotten. As long as we remember, Americans fought courageously in defence of liberty and justice regardless of how many times it took us to achieve victory! They did this while wearing red, white blue stripes and carrying flags decorated with stars! Do you realize that all of these events happened during your lifetime? ! That's great news, Your Highness.Now that you've had some extra time to reflect on the past let's discuss something less serious shall we?"

Figure 12: A case study of paraphrased text span detection. The upper part presents the paraphrased text while the lower part denotes the original text. The red underlined text represents the paraphrased text spans, and the orange background indicates model predictions. Darker colors indicate a higher degree of paraphrasing. The text in the blue background represents the original text span before paraphrasing.

Paraphrased Text

I hope that one of the airports I often visit will have a good Mexican eatery. But that is impossible. Short and sweet: Good: The steak quesadilla was alright. The tacos were delicious, however, the staff lacked friendliness and helpfulness. I plan to visit this place again in the future! So far I've had great experiences at this establishment and hope they keep up the good work. I really enjoy the shrimp burrito bowl the most. But what I'm most thankful for is that my son took me here for his first date with Kaylee. Actually, he wants to come back again because he really loves it, sweetheart! Finding a fresh dining option can occasionally become exasperating. We aim to explore unique restaurants each time we visit our place of origin. It wasn't until recently when after visiting two other places across town where neither had any food options that met my standards that we decided to come back again to Hickory Hill Brewery Restaurant Tasting Room which serves amazing food from start to finish and it's all made right there by talented chefscooks who make each dish unique using only locally sourced ingredients. What makes their food truly extraordinary? If you're expecting a typical American menu with items like burgers and chicken wings, you might be surprised. Instead, the menu showcases exceptional main course options, including seafood dishes like lobster bisque, avocado scallops, and lobster roll with creamy mashed potatoes. You'll also find tender steaks served with various pasta sauces, chili con carne, and mac 'n' cheese. One dish that stood out to us was the Brought Down My Baby Chicken Wings—truly delectable. Ordering dessert or drinks is not necessary, as even children can enjoy them while parents relax and enjoy some quality time together without the rush of service or hungry mouths waiting outside the dining room. Food wise no complaints whatsoever! Was let down however when I called to express concern over how long it took to order a ground beef sandwich due to me demanding to speak to someone immediately rather then having to wait till 5:30pm to tell a random server "I called earlier." Not being used to sitting around waiting we requested a table by the window but upon arriving didn't see anyone except maybe 3 people seated next to us giving off a vibe like we're pretty much alone in the place..

Raw Text

So I wish one of these airports I frequent would have a nice mexican restaurant. But that is impossible. Short and sweet: Good: The steak quesadilla was alright. The tacos were good, but the servers just weren't very friendly or helpful. I will be returning to this location in the future as well! So far I've had great experiences at this establishment and hope they keep up the good work. My favorite items are the shrimp burrito bowl, for sure! More than anything else though, I'm glad my son brought me here on his first date with Kaylee. In fact, he's planning another trip because he loves it so much baby girl! Trying to find something new to eat out can get frustrating sometimes.We try to go somewhere different every time we visit our hometown. It wasn't until recently when after visiting two other places across town where neither had any food options that met my standards that we decided to come back again to Hickory Hill Brewery Restaurant Tasting Room which serves amazing food from start to finish and it's all made right there by talented chefscooks who make each dish unique using only locally sourced ingredients. What makes their food truly extraordinary? It's not quite what you should expect if you're expecting a standard American menu featuring burgers, cheeseburgers and chicken wings (although those might also appear on an occasional night), instead you'll experience cooking techniques such as: The entire menu features outstanding main course choices including seafood dishes like lobster bisque, avocado scallops and lobster roll with creamy mashed potatoes along with tender steaks served over pasta sauces, chili con carne, mac n cheese and more. We found Brought Down My Baby Chicken Wings the best wing recipe we've ever tried - they're absolutely delicious!. Ordering dessert andor drinks really isn't needed, even kids enjoy them while parents sit down and hangout together enjoying some time away from home without needing to worry about rushed service or unsatisfied hungry mouths waiting patiently outside the dining room. Food wise no complaints whatsoever! Was let down however when I called to express concern over how long it took to order a ground beef sandwich due to me demanding to speak to someone immediately rather then having to wait till 5:30pm to tell a random server "I called earlier." Not being used to sitting around waiting we requested a table by the window but upon arriving didn't see anyone except maybe 3 people seated next to us giving off a vibe like we're pretty much alone in the place..

Figure 13: A case study of paraphrased text span detection. The upper part presents the paraphrased text while the lower part denotes the original text. The red underlined text represents the paraphrased text spans, and the orange background indicates model predictions. Darker colors indicate a higher degree of paraphrasing. The text in the blue background represents the original text span before paraphrasing.

Paraphrased Text

I am a local and went to Buddy V's for the first time. I was very unimpressed with the overall meal. The service was incredibly poor. Once we sat down they were exceptionally rude, having been served by the dude who had shoved his hands into the table causing him to jump up and down like he couldn't even use any of his own wine glasses at all thanks to his drunk stupidity (is that weird?) Then there wasn't much conversation so we left without saying anything because they were too busy taking pictures. After we finished our meal, the waiter insisted on contacting us immediately. When we asked why, we suggested that the waiter simply inform us it was time to leave instead. They explained that Buddy V, who is considered a local, should have better English skills than us. We promptly returned to the restaurant but were clueless about the situation until we witnessed Buddy V discussing the benefits of us surrendering control to him in his establishment. Despite this, the food was amazing. Later, the waiter came out and informed us that Buddy V had inquired about our satisfaction, but also mentioned that we should not eat since Buddy V lacked money and supplies. Nevertheless, we will definitely treat Buddy V to dinner the next time we visit. And next time Buddy V gives Buddy V dinner please implement a rule where you don't touch our food TRULY! : Buddy V is already dead set against eating our food! He said eating snacks makes him feel as bad as eating pizza! He also tried to make Buddy V snack for his friends once before they got together!!!!!!!!!! It goes both ways!!! Good luck Buddy V!!!! ;D Louise

Raw Text

I am a local and went to Buddy V's for the first time. I was very unimpressed with the overall meal. The service was incredibly poor. Once we sat down they were exceptionally rude, having been served by the dude who had shoved his hands into the table causing him to jump up and down like he couldn't even use any of his own wine glasses at all thanks to his drunk stupidity (is that weird?) Then there wasn't much conversation so we left without saying anything because they were too busy taking pictures. When we finished eating, the waiter insisted on calling us back immediately after. After asking what happened we asked why the waiter didn't just tell us it was our turn to leave instead and they explained that Buddy V is our "local" which meant he needed to speak English better than us! We arrived right away but we had no idea what was happening until we saw Buddy V talking about how much better life would be if we gave him free reign over us in his restaurant!!! The food was awesome though!! The waiter came out and told us Buddy V asked if everything was okay, then proceeded to tell us we should not eat since Buddy V doesn't have money or supplies!!!! We will definitely get Buddy V dinner again next time around! ! And next time Buddy V gives Buddy V dinner please implement a rule where you don't touch our food TRULY! : Buddy V is already dead set against eating our food! He said eating snacks makes him feel as bad as eating pizza! He also tried to make Buddy V snack for his friends once before they got together!!!!!!!!!! It goes both ways!!! Good luck Buddy V!!!! ;D Louise

Figure 14: A case study of paraphrased text span detection. The upper part presents the paraphrased text while the lower part denotes the original text. The red underlined text represents the paraphrased text spans, and the orange background indicates model predictions. Darker colors indicate a higher degree of paraphrasing. The text in the blue background represents the original text span before paraphrasing.

	Sentence Reordering
Raw Text 1	It felt good to know that his efforts were appreciated. At the end of the week, Paul was exhausted but satisfied.
Paraphrase 1	At the end of the week, Paul was exhausted, but satisfied. It felt good to know that his efforts were appreciated.
Raw Text 2	Its happened a few times, like some gas leak in some town and it instructed for people to be indoors and shut all the windows. Signal intrusion has happened before as someone already posted. But encryption makes that all but impossible today. You would either need some serious ability to break encryption or infiltrate multiple tv providers to pull it off.
Paraphrase 2	Signal intrusion has happened before as someone already posted. But encryption makes that all but impossible today. You would either need some serious ability to break encryption or infiltrate multiple TV broadcasters to pull it off. It's happened a few times, like some gas leak in some town and it instructed for people to be indoors and shut all the windows.
Raw Text 3	Keep the toes of your forward foot pointing upwards. To choose which leg to bend, try both out and see which one feels most comfortable to you. Determine how far apart the legs should be when sitting on a chair or stool: If they are together, then position yourself so that your heels touch.
Paraphrase 3	Determine how far apart the legs should be when sitting on a chair or a stool: if they are together, then position yourself so that your heels touch. Keep the toes of the front foot pointing upwards. To choose which leg to bend, try both legs and see which one is more comfortable to you.
	Word Replacement
Raw Text 4	Among those arrested were six suspects in Italy, four in Britain, and three in Norway. Police say some of the suspects may have travelled to Syria or Iraq. Italy's Ansa Six suspected Ansar al-Sharia fighters were arrested in Italy, Britain, Pakistan, Norway, Scotland and Germany, bringing the total to 18. Police say they have arrested a further 22 people on suspicion of involvement in terrorism.
Paraphrase 4	Among those arrested were six suspects in the United States, four in Britain, and three in Norway. Police say some of the suspects may have travelled to Syria or Iraq. Italy's Ansa Six suspected Jaish al-Sharia fighters were arrested in Italy, Britain, Pakistan, Norway, France and Germany, bringing and the total to 181. Police say they have arrested a further 22 people on suspicion of involvement in terrorism.
Raw Text 5	Soviet Union. Russia was a state of the Soviet Union. It technically took the Germans a while to reach Russia after they invaded.
Paraphrase 5	Soviet Union. Russia was a state of the Soviet Union. It also took the Germans a while to conquer Russia after they invaded.

Table 8: Case illustration of two types of minor modifications: sentence reordering and word replacement.