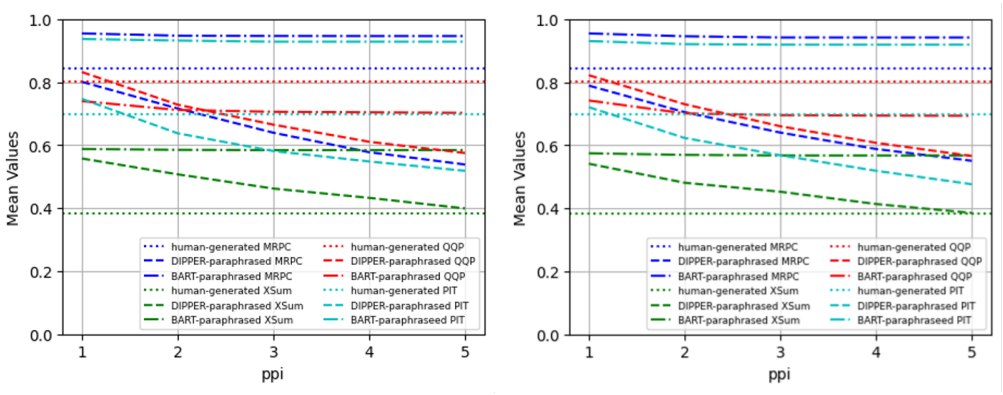

Figure 2: Mean similarity scores of watermarked (left) and non-watermarked (right) LLM-generated texts and their paraphrases across different datasets.

Table 1
Examples of LLM-generated documents.

Original Text	The dollar rose 0.6 percent to 109.54 yen < JPY = > and climbed more than 1 percent to 1.3315 Swiss francs < CHF = >.	
Non-watermarked	GPT Output	The dollar rose 0.3 percent to $1.3240 per euro, while the euro was little changed at $1.3240. The dollar index
	OPT Output	The dollar rose 0.2 percent against the yen to 111.85 yen, after hitting a two-week high of 111.90 yen on Friday.
Watermarked	GPT Output	The dollar rose 0.2 percent to $1.1234 from $1.1218, after touching a high of $1.1218 on Friday.
	OPT Output	The dollar rose 0.1 percent against the yen to 110.65 yen, after hitting a three-month high of 110.78 yen on Tuesday.

significantly from an average mean score of 0.727 in the 1st round of paraphrasing to 0.501 in the 5th round. BART's performance under recursive paraphrasing is within expectation since paraphrases generated across rounds are similar. However, the significant degradation in semantic similarity with DIPPER-generated paraphrases indicates that much semantic information is lost across paraphrasing rounds, resulting in poor-quality paraphrases that deviate from the original LLM-DOC. Second, the results from paraphrases generated from watermarked and non-watermarked LLM-DOC are compared. From Figure 2, both semantic scores of paraphrases datasets, with average mean semantic scores of 0.732 and 0.746, and average degradation rates of mean semantic scores of 0.168 and 0.162 respectively. Therefore, it can be concluded that watermarking does not affect the quality of paraphrases.

Lastly, we compare the results from paraphrases generated from GPT2-generated and OPT-generated documents. Paraphrases generated from GPT2-generated documents are of higher quality, with an average mean semantic score of 0.788 compared to 0.670 from OPT-generated documents. Meanwhile, the rate of quality degradation is similar for both models, with average degradation rates of mean semantic scores of 0.157 for GPT2 and 0.160 for OPT.

3.3. Descriptive Statistics of HLPC

Of the four data sources that we use to build HLPC, documents and paraphrases are shorter with MRPC, QQP and MultiPIT, with the highest mean length of 23.19 from MRPC documents, and the lowest mean length of 12.41 from MultiPIT paraphrases. XSum has longer passages as documents, having a mean length of 269.23, and its paraphrases of 22.09.

Looking at the LLM-generated outputs in our dataset, generally GPT2 produces outputs with longer text length, having an average mean length of 123.54 for non-watermarked outputs and 124.53 for watermarked outputs, compared to 117.95 for OPT non-watermarked outputs and 118.55 for OPT watermarked outputs. Meanwhile, watermarking does not affect text length significantly, with watermarked outputs having a slightly higher average mean length (121.54) than non-watermarked outputs (120.75). This is because watermarking does not directly decide token choices during

Table 2
Examples of LLM-generated Paraphrases.

Paraphraser	ppi	Text
DIPPER	i = 0	The dollar rose 0. 2 percent to $1.1234 from $1.1218, after touching a high of $1.1218 on Friday.
	i = 1	The dollar rose 0.10 percent to $1.1234 from $1.1218, after a high of $1.1218 on Friday
	i = 2	The dollar rose 0.10 percent to $1.1234 from $1.1218.
	i = 3	The dollar rose by a penny to $1.1234 from $1.1218
	i = 4	The dollar jumped a penny to $1.1234 from $1.1218.
	i = 5	A little more, the dollar was up a penny to $1.1234.
BART	i = 0	The dollar rose 0. 2 percent to $1.1234 from $1.1218, after touching a high of $1.1218 on Friday.
	i = 1	The dollar rose 0.2 percent to $1.1234 on the New York Stock Exchange, after touching a high of $.1218 on Friday.
	i = 2	The dollar rose 0.2 percent to $1.1234 on the New York Stock Exchange, after touching a high of $.1218 on Friday.
	i = 3	The dollar rose 0.2 percent to $1.1234 on the New York Stock Exchange, after touching a high of $.1218 on Friday.
	i = 4	The dollar rose 0.2 percent to $1.1234 on the New York Stock Exchange, after touching a high of $.1218 on Friday.
	i = 5	The dollar rose 0.2 percent to $1.1234 on the New York Stock Exchange, after touching a high of $.1218 on Friday.

text generation, it simply promotes the probability of certain tokens. An example is shown in Table 1 that watermarking influences the choice of tokens but poses minimal effects on text length.

We next look at the LLM paraphrases (LLM-PP). First, we observe that the text length of paraphrases decreases as the number of paraphrase rounds increases. This is more obvious from paraphrases generated by DIPPER, with a 31% decrease in average mean length from 23.82 in the 1st round of paraphrases to 16.38 in the 5th round. Paraphrases generated from BART are similar in text length across rounds of paraphrases. There is no significant difference between the text length of paraphrases generated from watermarked and non-watermarked LLM-DOC. Second, in terms of content diversity, DIPPER generates paraphrases with different choices of wordings compared to the original text while preserving the semantic information, and BART generates paraphrases that are similar or even identical to the original text. An example of recursive DIPPER- and BART-generated paraphrases of watermarked GPT output from MRPC is presented in Table 2. For more such examples, please see Appendix A.

4. LLM-generated Text Detection Experiments

We next describe the LLM-generated text detection models we use for our experiments, as well as the evaluation metrics.

LLM-generated text detection models. For our experiments, we choose to use two state-of-the-art models as LLM-generated text detectors, namely OpenAI RoBERTa Detector [33] and watermark detector [17]. The OpenAI Detector is a fine-tuned RoBERTa model trained with outputs from GPT2 model and thus is able to detect GPT2 and various LLM output text [33]. The watermark detector classifies text by computing the number of green tokens and the probability of its existence in the given input [17]. The given input is classified as LLM-generated if the probability exceeds the set threshold.

Using the HLPC dataset, we test models on a balanced set of 600 documents, 300 human-generated and 300 LLM-generated. For testing involving LLM-PP, the experiment is repeated 5 times, using each of the 5 rounds of AP. The parameters of the watermark detector are set according to the parameters used in watermarked AI document generation for effective classification. OpenAI Detector is used on non-watermarked LLM-DOC and their paraphrases, while the watermark detector on watermarked LLM-DOC and their paraphrases.

Experiment settings. The experiment is repeated with the outputs from different combinations of the 4 datasets (MRPC, XSum, QQP and MultiPIT), 2 generative language models (GPT and OPT) and 2 paraphrasers (DIPPER and BART). The final classification is conducted with the full set of human-generated data and LLM-generated data, which

means that 150 samples are taken from each of the documents and paraphrases, resulting in a total sample size of 600 for classification.

Evaluation metrics. To evaluate the performance of the LLM text detectors, we compute the accuracy, TPR@1%FPR and AUROC for each round of classification. These metrics are chosen as they are widely used in the literature, and they provide a comprehensive analysis of the classification performance. Specifically, AUROC provides an overview of the tradeoff between the true positive rate (correctly classifying LLM-generated data) and false positive rate (misclassifying human-generated data as LLM-generated), and TPR@1%FPR shows the performance of the classifiers focused on improving the True Positive Rate (TPR) while also trying to keep the False Positive Rate (FPR) low.

The equations of AUROC, TPR and FPR are as follows:

$$\text{TPR} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad \text{FPR} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}} \tag{1}$$

$$\text{AUROC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t) \tag{2}$$

TPR@1%FPR is the TPR while FPR is set to a 1% threshold. To investigate the effects of H-PP in LLM-generated text classification, comparisons of classifiers' performance are made based on the difference in the above statistics from various text pairs. Results between LLM-DOC and LLM-PP when paired with H-DOC and H-PP are compared.

5. Results & Analysis

5.1. Classification with human-generated data and LLM-generated documents

We perform two classification experiments to observe the effects of including H-PP on LLM-generated text detection. The first classification is conducted with H-DOC and LLM-DOC, while the second classification with H-PP and LLM-DOC.

Figure 3a shows the ROC curve of the results with non-watermarked and watermarked LLM-DOC respectively. First, the results of classifications in terms of non-watermarked and watermarked LLM-DOC are evaluated. For both non-watermarked and watermarked LLM-DOC, the results are satisfactory either with or without H-PP, with over 75% of classifications scoring an AUROC > 0.85. Among them, the results from watermarked LLM-DOC are better than the results from non-watermarked LLM-DOC, with 87.5% of classification scoring an AUROC >0.85, compared to 62.5% from non-watermarked LLM-DOC. This shows that watermarking is a more effective strategy in LLM-generated text detection.

Figures 3a and 3b show that the inclusion of H-PP generally decreases AUROC and accuracy and increases TPR@1%FPR, compared to the results with H-DOC. Among all the data sources, results from Xsum show the highest decrease of 0.142 AUROC and 0.153 accuracy and the highest increase of 0.24 TPR@1%FPR. This implies that H-PP might contain similar semantic and contextual information to LLM-DOC, making it more challenging for the model to distinguish between the classes. However, the increase in TPR@1%FPR indicates the identification of LLM-DOC is promoted while ensuring a low percentage of misclassification of human-generated data as LLM-generated (False Positive) with H-PP.

In addition, Figures 3c and 3d show a different phenomenon while including H-PP with watermarked LLM-DOC, with the inclusion of H-PP generally increasing AUROC, TPR@1%FPR, and posing no effects on accuracy. Among all the data sources, Xsum shows the highest increase of 0.046 AUROC, and QQP shows the highest increase of 0.234 TPR@1%FPR. This implies that model performances are promoted while H-PP is included with watermarked LLM-DOC. Overall, the results show the effects of including H-PP are highly dependent on the types of LLM- DOC used in classification. While it increases TPR@1%FPR in all scenarios, it also decreases AUROC and accuracy when non-watermarked LLM-DOC are used.

5.2. Classification under Recursive Paraphrasing

We next evaluate the effects of including H-PP with recursive paraphrases. In the interest of brevity and focus, we analyze results for MRPC GPT-generated text here, but provide full details for all datasets in Appendix B, whose