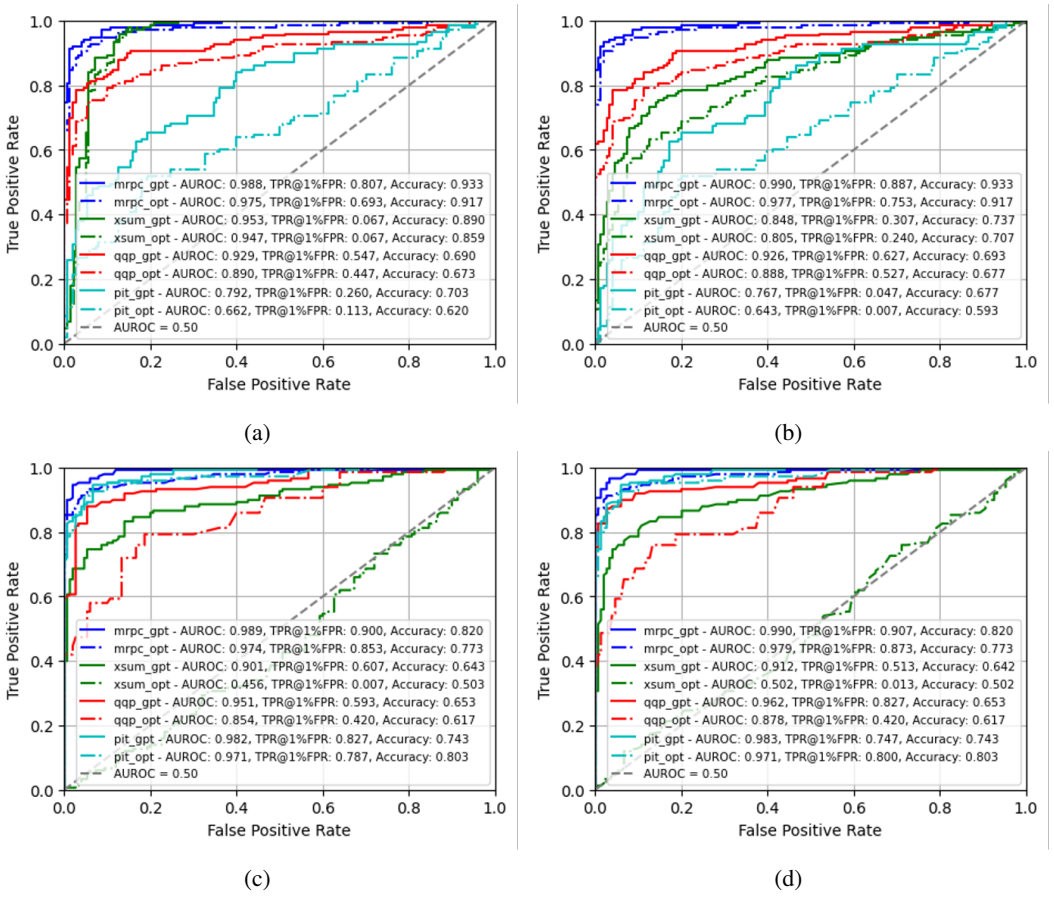

Figure 3: ROC Curve Comparison: (a) Human-generated Documents vs Non-watermarked LLM-generated Documents, (b) Human-generated Paraphrases vs Non-watermarked LLM-generated Documents, (c) Human-generated Documents vs Watermarked LLM-generated Documents, (d) Human-generated Paraphrases vs Watermark LLM-generated Documents

results are consistent with MRPC. The results (H-PP_pp0, H-PP_pp2 & H-PP_pp5) are compared to the classification result without the use of H-PP (pp0, pp2 & pp5). First, the general model performance is discussed in terms of the differences between DIPPER and BART-generated paraphrases.

Figure 4 shows that regardless of the presence of watermarking and H-PP, model performances with DIPPER-generated paraphrases degrade to a larger extent than with BART-generated paraphrases across rounds of paraphrasing. A decrease of 0.26 in average AUROC and 0.67 in average TPR@1%FPR are observed from DIPPER, while BART's average AUROC and TPR@1%FPR decrease merely 0.0075 and 0.045 respectively. Particularly, the classification with DIPPER-generated paraphrases from watermarked LLM-DOC (Figure 4c) shows the highest degradation in model performance, resulting in the lowest AUROC of 0.49 which is worse than a random classifier. This can be attributed to the lower semantic similarity between the LLM-DOC and DIPPER-generated LLM-PP mentioned in Section 3.2. The low semantic similarity might indicate that DIPPER-generated LLM-PP becomes more similar to H-PP coincidentally while deviating from the original LLM-DOC, resulting in a degradation in model performance.

In addition, classification with BART-generated LLM-PP (Figure 4d) shows excellent results with AUROC $\tilde{0}$.98 even after 5 rounds of paraphrasing. Second, the effects of including H-PP in classification are evaluated. In both classifications with paraphrases from watermarked and non-watermarked LLM-DOC, the performances of the models with H-PP are better than those without H-PP, while the effect of the promotion is more significant with paraphrases generated from non-watermarked LLM-DOC. For classification with paraphrases generated from non-watermarked LLM-DOC and H-PP (Figures 4a and 4b), TPR@1%FPR increases from a minimum of 0.02 to a maximum of 0.153 when compared to classification with H-DOC. The positive effect is less significant on AUROC with an average increase of 0.0028. Although minimal effects are cast on AUROC and accuracy with the use of H-PP, the significant

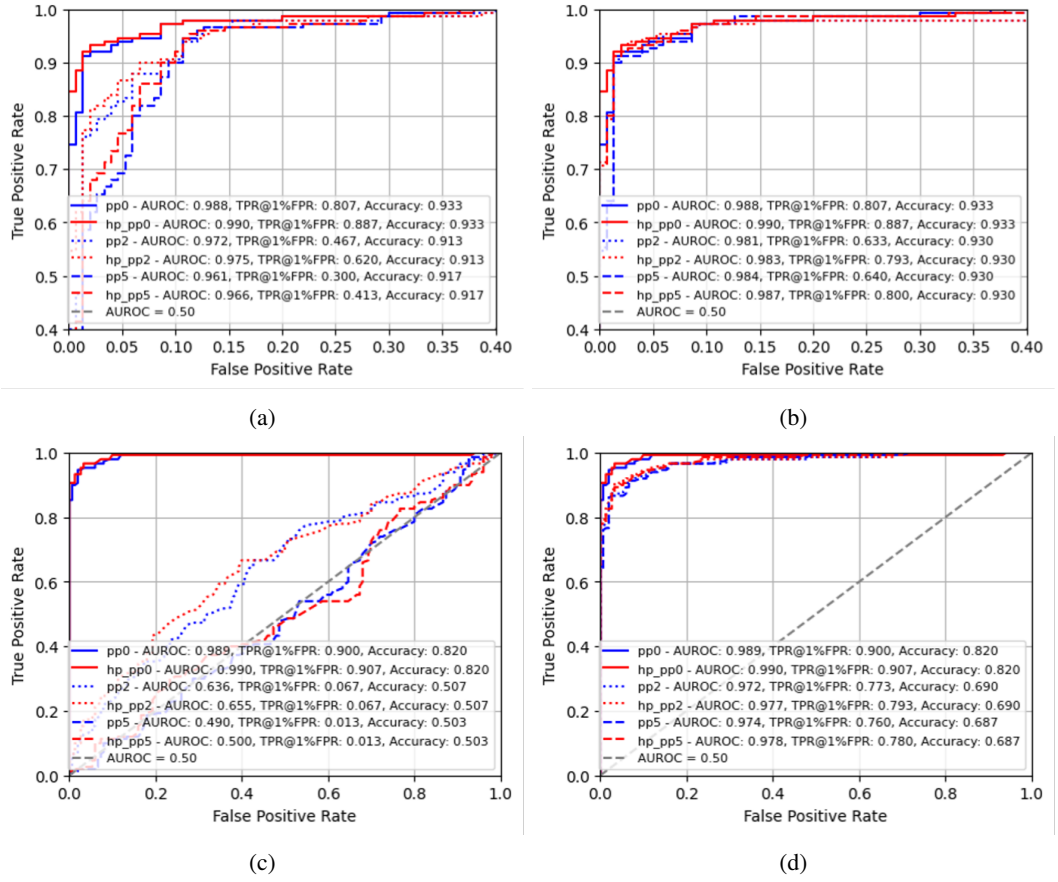

Figure 4: ROC Curve Comparison: (a) Human-generated Documents/Paraphrases vs DIPPER-generated Paraphrases from Non-watermarked MRPC GPT2-generated Text, (b) Human-generated Documents/Paraphrases vs BART-generated Paraphrases from Non-watermarked MRPC GPT2-generated Text, (c) Human-generated Documents/Paraphrases vs DIPPER-generated Paraphrases from Watermarked MRPC GPT2-generated Text, (d) Human-generated Documents/Paraphrases vs BART-generated Paraphrases from Watermarked MRPC GPT2-generated Text

increase in TPR@1%FPR shows that H-PP helps in the detection of LLM-generated data at a low false positive rate, ensuring that less human data are classed as LLM-generated.

For classification with paraphrases generated from watermarked LLM-DOC and H-PP (Figures 4c and 4d), minimal increases in AUROC and TPR@1%FPR are observed, with the highest increase in AUROC of 0.019 and TPR@1%FPR of 0.02. Overall, it can be concluded that including H-PP in the classification under recursive paraphrasing helps promoting AUROC and TPR@1%FPR under recursive paraphrasing. The comparison of classification results between the inclusion and exclusion of H-PP with other datasets and generative models are also presented in Appendix B. Generally, similar results are observed, except for classification with Xsum where the inclusion of H-PP reduces AUROC.

5.3. Classification with Full Set of Human-generated Data and LLM-generated Data

We now perform classification experiments with the full set of human-written and LLM-generated data, passing both documents and paraphrases to the classifier.

Table 3 shows the comparison of the average results from classification with i) H-DOC vs LLM-PP (pp1 & pp5), ii) H-PP vs LLM-PP (H-PP_pp1 & H-PP_pp5) and iii) full human data vs LLM-generated data (f_pp1 & f_pp5) across all datasets. First, for results from paraphrases generated from watermarked LLM-DOC, model performance shows extreme opposites depending on the paraphrasers used. The best improvement in statistical results is shown using DIPPER-generated paraphrases from watermarked LLM-DOC, along with H-DOC, H-PP and LLM-DOC (f_pp1 &

Table 3
Average results of classification with full human-written and LLM-generated data; bracketed numbers indicate percentage difference.

LLM-generated Documents	Paraphraser	Data	AUROC	TPR1%FPR	Accuracy
Non-watermarked	DIPPER	pp1	**0.908**	0.285	**0.814**
		hp_pp1	0.883	0.358	0.769
		f_pp1	0.884 (-2.71%)	**0.379 (32.98%, 5.86%)**	0.778 (-4.62%)
		pp5	**0.878**	0.138	**0.811**
		hp_pp5	0.843	0.169	0.766
		f_pp5	0.867 (-1.27%)	**0.290 (110.14%, 71.60%)**	0.775 (-2.32%)
	BART	pp1	0.873	0.209	0.793
		hp_pp1	0.839	0.284	0.750
		f_pp1	**0.874 (0.001%, 4.17%)**	**0.313 (49.76%, 10.21%)**	**0.771 (-2.85%)**
		pp5	**0.876**	0.201	**0.795**
		hp_pp5	0.842	0.276	0.751
		f_pp5	0.874 (-0.002%)	**0.341 (69.65%, 23.55%)**	0.773 (-2.85%)
Watermarked	DIPPER	pp1	0.652	0.100	0.508
		hp_pp1	**0.678**	0.117	0.508
		f_pp1	0.636 (-6.60%)	**0.209 (109%, 78.63%)**	**0.556 (9.45%, 9.45%)**
		pp5	0.512	0.019	0.500
		hp_pp5	0.541	0.022	0.500
		f_pp5	**0.587 (14.65%, 8.5%)**	**0.185 (873.68%, 740.91%)**	**0.554 (10.8%, 10.8%)**
	BART	pp1	0.825	0.445	**0.611**
		hp_pp1	**0.840**	**0.467**	**0.611**
		f_pp1	0.706 (-18.98%)	0.315 (-48.25%)	0.584 (-4.62%)
		pp5	0.797	0.398	**0.599**
		hp_pp5	**0.813**	**0.426**	0.598
		f_pp5	0.700 (-16.14%)	0.307 (-38.76%)	0.581 (-3.10%)

f_pp5). With the 1st round of paraphrases, TPR@1%FPR increases by 109% and 78.63% and accuracy increases by 9.45% and 9.45%, compared to only using H-DOC or H-PP respectively. With the 5th round of paraphrases, AUROC increases by 14.65% and 8.5%, TPR@1%FPR increases by 874% and 740% and accuracy increases by 10.8% and 10.8%, compared to only using H-DOC or H-PP respectively. This shows that using the full set of data as input for classification is significantly effective in improving LLM-generated text detection under recursive paraphrasing, with the condition that the LLM-DOC is watermarked and paraphrases are DIPPER-generated. Meanwhile, the worst performance is shown with BART-generated paraphrases from watermarked LLM-DOC under the same condition of using the full set of data as input. Compared with the best results, results from 1st and 5th rounds of BART-generated paraphrases show a degradation of 18.98% and 16.14% in AUROC, 48.25% and 38.76% in TPR@1%FPR and 4.62% and 3.10% in accuracy respectively. The significant difference in the statistical results shows that the watermark detector is highly sensitive to the paraphraser used. While the performance improves significantly with DIPPER-generated paraphrases, it also degrades significantly with BART-generated paraphrases.

Second, TPR@1%FPR significantly increases while using paraphrases generated from non-watermarked LLM-DOC along with H-DOC, H-PP and LLM-DOC, while AUROC and accuracy remain unchanged or decrease slightly. Results show that TPR@1%FPR increases by a minimum of 5.86% to a maximum of 110.14%, with an average of 42.84% across different inputs. This indicates that having the full set of data as input effectively improves TPR@1%FPR and ensures minimal misclassification at a low false positive rate. Meanwhile, AUROC and accuracy decrease, compared to the results from H-DOC vs LLM-PP and H-PP vs LLM-PP. However, the decrease is less significant compared to the increase in TPR@1%FPR. The maximum decrease in AUROC and accuracy is merely 4.17% and 4.62% respectively. Therefore, it can be concluded that having a full set of data as input exhibits a trade-off between accuracy, AUROC and TPR@1%FPR. Considering the significant improvement in TPR@1%FPR and the importance of ensuring minimal misclassification, having a full set of data as input is suggested to be a better method than using