merely H-DOC under recursive paraphrasing. Detailed ROC curve and statistical results of classification with each dataset and its generations are presented in Appendix C.

5.4. Benefits of Using Human-written Paraphrases in LLM or Detectors Training

In our review of the literature in Section 2 we found that existing LLMs are trained on corpora that do not contain H-PP information. Since existing detectors are designed to identify the LLM-generated statistical pattern and watermark from the input text, paraphrasing which reduces or erases the above characteristics could effectively evade the detectors. Our experiments show that including H-PP in the dataset promotes classification performances under different circumstances, and therefore, including H-PP in the training datasets during the training of detectors could effectively improve detectors' classification performances since models could learn about the fundamental differences of semantic and contextual information between human-written and LLM-generated text, even under recursive paraphrasing.

Our results also show that the effectiveness of including H-PP in the dataset is highly dependent on the existence of watermarking and the type of paraphraser used. In our experiments, while H-DOC is included for classification, watermarking and DIPPER-generated paraphrases help improve classification performance, while experiments with non-watermarked and BART-generated paraphrases show the opposite. As such, it is important to understand and predict the potential usage of watermarking and the type of paraphraser while developing the detectors. Detector developers could either get the information from users or employ a multi-step classification model for accurate prediction. A multi-step classification model could first identify the presence of watermark and the type of paraphrasers, then decide whether to include H-PP in the training dataset of the detectors based on the results. If such technology or information is not available, it is recommended to include either H-DOC or H-PP, to avoid significant degradation in classification performance. Meanwhile, we show that including H-PP in the datasets is highly effective under recursive paraphrasing. As such, we recommend that detectors, which are used in circumstances where paraphrasing is prevalent, for example, in academic publications, to be trained with H-PP instead of only H-DOC, so as to increase AUROC and TPR@1%FPR.

6. Conclusion

In this study, our aim was to investigate the effect of human paraphrases (H-PP) on LLM-generated text detection by conducting classifications with various combinations of human and LLM-generated data pairs. To enable this study, we devise a data collection strategy and generate the HLPC dataset by leveraging and extending four existing data sources: MRPC, XSum, QQP and MultiPIT. Unlike previous datasets, our new dataset, Human & LLM Paraphrase Collection (HLPC), incorporates human-written documents (H-DOC), human-written paraphrases (H-PP), LLM-generated texts (LLM-DOC) and LLM-generated paraphrases (LLM-PP). We generate LLM documents by prompting GPT2-XL and OPT-13B with prompts derived from human-written documents. AI paraphraser, DIPPER and BART are then used to paraphrase the generated outputs. Using this dataset, we perform classification experiments with state-of-the-art LLM-generated text detectors OpenAI RoBERTa and watermark detector, with the aim of understanding the effects of incorporating human-written paraphrases in LLM-generated text detection. Data pairs used for classifications include i) H-DOC vs LLM-DOC, ii) H-PP vs LLM-DOC, iii) H-DOC vs LLM-PP, iv) H-PP vs LLM-PP and v) H-DOC & H-PP vs LLM-DOC & LLM-DOC. 3 comparisons are made between the classification results to examine the effects of including H-PP in classification. First, results from (i) and (ii) are compared to show H-PP's effects while classification is done with LLM-DOC. Second, results from (iii) and (iv) are compared to show H-PP's effects under recursive paraphrasing. Lastly, results from (v) are compared to results from (iii) and (iv) to examine the effects of having a full set of human and LLM-generated data.

In our experiments, we observe that in all 3 sets of comparisons, including H-PP in the classification is effective in promoting TPR@1%FPR, while its effects on AUROC and accuracy are highly dependent on the presence of watermarking and the type of paraphraser. In the 1st set of comparison, the results show that TPR@1%FPR increases in all scenarios, but AUROC and accuracy decrease if non-watermarked LLM-DOC are used. For the 2nd set of comparison, AUROC and TPR@1%FPR increases to a small extent in all scenarios, while accuracy remains unchanged under recursive paraphrasing. Lastly, for the 3rd set of comparison with the full set of data, results vary in 2 extremes depending on the paraphraser used to generate paraphrases from watermarked LLM-DOC, while TPR@1%FPR increase significantly and AUROC and accuracy decrease slightly with non-watermarked LLM-DOC and their paraphrases. Therefore, it can be concluded that the inclusion of H-PP in classification promotes TPR@1%FPR with a possible trade- off of AUROC and accuracy.

Our study has potential to be further extended in the future by studying additional datasets and detection models, to tackle some of the limitations of our study. First, the sentences in the chosen datasets are relatively short, with a mean length of 48.68 tokens. Since the performance of LLM-generated text detectors increases with the input text length, consideration of additional datasets with longer sentences would help provide a more diverse analysis of the effects of H-PP's inclusion in classification. However, due to the limited availability of datasets that contain H-PP, only datasets with short sentences are used in this project. Second, other state-of-the-art LLM text detection tools could be tested to broaden the findings, such as GPTZero,[2] which was excluded from our study due to the associated costs.

References

[1] Allam, A.M.N., Haggag, M.H., 2012. The question answering systems: A survey. International Journal of Research and Reviews in Information Sciences (IJRRIS) 2.

[2] An, H., Acquaye, C., Wang, C., Li, Z., Rudinger, R., 2024. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? arXiv preprint arXiv:2406.10486 .

[3] Ansari, S., 2023. sentence_similarity_semantic_search. URL: https://github.com/Sakil786/sentence_similarity_semantic_search. gitHub repository.

[4] Barreto, F., Moharkar, L., Shirodkar, M., Sarode, V., Gonsalves, S., Johns, A., 2023. Generative artificial intelligence: Opportunities and challenges of large language models, in: International Conference on Intelligent Computing and Networking, Springer. pp. 545–553.

[5] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA. p. 610–623. URL: https://doi.org/10.1145/3442188.3445922, doi:10.1145/3442188.3445922.

[6] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. URL: https://arxiv.org/abs/2005.14165, arXiv:2005.14165.

[7] Damodaran, P., 2021. Parrot: Paraphrase generation for nlu.

[8] Dictionary, C., 2019. Paraphrase | meaning in the cambridge english dictionary. URL: https://www.cambridge.org/dictionary/english/paraphrase.

[9] Dolan, B., Brockett, C., 2005. Automatically constructing a corpus of sentential paraphrases, in: Third International Workshop on Paraphrasing (IWP2005). URL: https://www.microsoft.com/en-us/research/publication/automatically-constructing-a-corpus-of-sentential-paraphrases/.

[10] Dou, Y., Jiang, C., Xu, W., 2022. Improving large-scale paraphrase acquisition and generation. URL: https://arxiv.org/abs/2210.03235, arXiv:2210.03235.

[11] El-Kassas, W.S., Salama, C.R., Rafea, A.A., Mohamed, H.K., 2021. Automatic text summarization: A comprehensive survey. Expert systems with applications 165, 113679.

[12] Fröhling, L., Zubiaga, A., 2021. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. PeerJ Computer Science 7, e443.

[13] Goyal, R., Kumar, P., Singh, V.P., 2023. A systematic survey on automated text generation tools and techniques: application, evaluation, and challenges. Multimedia Tools Appl. 82, 43089–43144. URL: https://doi.org/10.1007/s11042-023-15224-0, doi:10.1007/s11042-023-15224-0.

[14] Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., Denuyl, S., 2020. Social biases in nlp models as barriers for persons with disabilities. URL: https://arxiv.org/abs/2005.00813, arXiv:2005.00813.

[15] Illia, L., Colleoni, E., Zyglidopoulos, S., 2023. Ethical implications of text generation in the age of artificial intelligence. Business Ethics, the Environment & Responsibility 32, 201–210. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/beer.12479, doi:https://doi.org/10.1111/beer.12479, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/beer.12479.

[16] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T., 2023. A watermark for large language models, in: International Conference on Machine Learning, PMLR. pp. 17061–17084.

[17] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T., 2024. A watermark for large language models. URL: https://arxiv.org/abs/2301.10226, arXiv:2301.10226.

[18] Kreps, S., McCain, R.M., Brundage, M., 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. Journal of Experimental Political Science 9, 104–117. doi:10.1017/XPS.2020.37.

[19] Krishna, K., Song, Y., Karpinska, M., Wieting, J., Iyyer, M., 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. URL: https://arxiv.org/abs/2303.13408, arXiv:2303.13408.

[20] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. URL: https://arxiv.org/abs/1910.13461, arXiv:1910.13461.

[21] Li, Z., Xu, X., Shen, T., Xu, C., Gu, J.C., Tao, C., 2024. Leveraging large language models for nlg evaluation: A survey. arXiv preprint arXiv:2401.07103 .

[22] Lin, S., Hilton, J., Evans, O., 2022. Truthfulqa: Measuring how models mimic human falsehoods. URL: https://arxiv.org/abs/2109.07958, arXiv:2109.07958.

[2] https://gptzero.me/

[23] Lopez, A., 2008. Statistical machine translation. ACM Computing Surveys (CSUR) 40, 1–49.
[24] Mishra, A.K., 2020. abhimishra91/transformers-tutorials. URL: https://huggingface.co/mrm8488/t5-base-finetuned-summarize-news.
[25] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C., 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. URL: https://arxiv.org/abs/2301.11305, arXiv:2301.11305.
[26] Mou, L., 2022. Search and learning for unsupervised text generation. AI Magazine 43, 344–352. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12068, doi:https://doi.org/10.1002/aaai.12068.
[27] Narayan, S., Cohen, S.B., Lapata, M., 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. URL: https://arxiv.org/abs/1808.08745, arXiv:1808.08745.
[28] Project, P.R.C.J., 2016. Reddit news users more likely to be male, young and digital in their news preferences. URL: https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-pre
[29] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2022. Robust speech recognition via large-scale weak supervision. URL: https://arxiv.org/abs/2212.04356, arXiv:2212.04356.
[30] Razaq, A., Halim, Z., Rahman, A., Sikandar, K., 2024. Identification of paraphrased text in research articles through improved embeddings and fine-tuned bert model. Multimedia Tools and Applications 83. doi:10.1007/s11042-024-18359-w.
[31] Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S., 2024. Can ai-generated text be reliably detected? URL: https://arxiv.org/abs/2303.11156, arXiv:2303.11156.
[32] Shewale, R., 2023. 30+ google bard statistics 2023 (trends & demographics). URL: https://www.demandsage.com/google-bard-statistics/.
[33] Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J.W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., Wang, J., 2019. Release strategies and the social impacts of language models. URL: https://arxiv.org/abs/1908.09203, arXiv:1908.09203.
[34] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T., 2023. Llama 2: Open foundation and fine-tuned chat models. URL: https://arxiv.org/abs/2307.09288, arXiv:2307.09288.
[35] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R., 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. URL: https://arxiv.org/abs/1804.07461, arXiv:1804.07461.
[36] Wolff, M., Wolff, S., 2022. Attacking neural text detectors. URL: https://arxiv.org/abs/2002.11768, arXiv:2002.11768.
[37] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., Choi, Y., 2020. Defending against neural fake news. URL: https://arxiv.org/abs/1905.12616, arXiv:1905.12616.
[38] Zhang, B., Haddow, B., Birch, A., 2023. Prompting large language model for machine translation: A case study, in: International Conference on Machine Learning, PMLR. pp. 41092–41110.
[39] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L., 2022. Opt: Open pre-trained transformer language models. URL: https://arxiv.org/abs/2205.01068, arXiv:2205.01068.
[40] Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., Hashimoto, T.B., 2024. Benchmarking large language models for news summarization. Transactions of the Association for Computational Linguistics 12, 39–57.
[41] Zhao, J., Chen, K., Yuan, X., Qi, Y., Zhang, W., Yu, N., 2024. Silent guardian: Protecting text from malicious exploitation by large language models. IEEE Transactions on Information Forensics and Security .
[42] Zubiaga, A., 2024. Natural language processing in the era of large language models. Frontiers in Artificial Intelligence 6, 1350306.

A. Examples of LLM-generated Paraphrases

Table 4: Examples of LLM-generated paraphrases; ppi means the i-th round of paraphrase.

MRPC

Paraphraser	Input	ppi	Text
	Watermarked GPT Output	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.