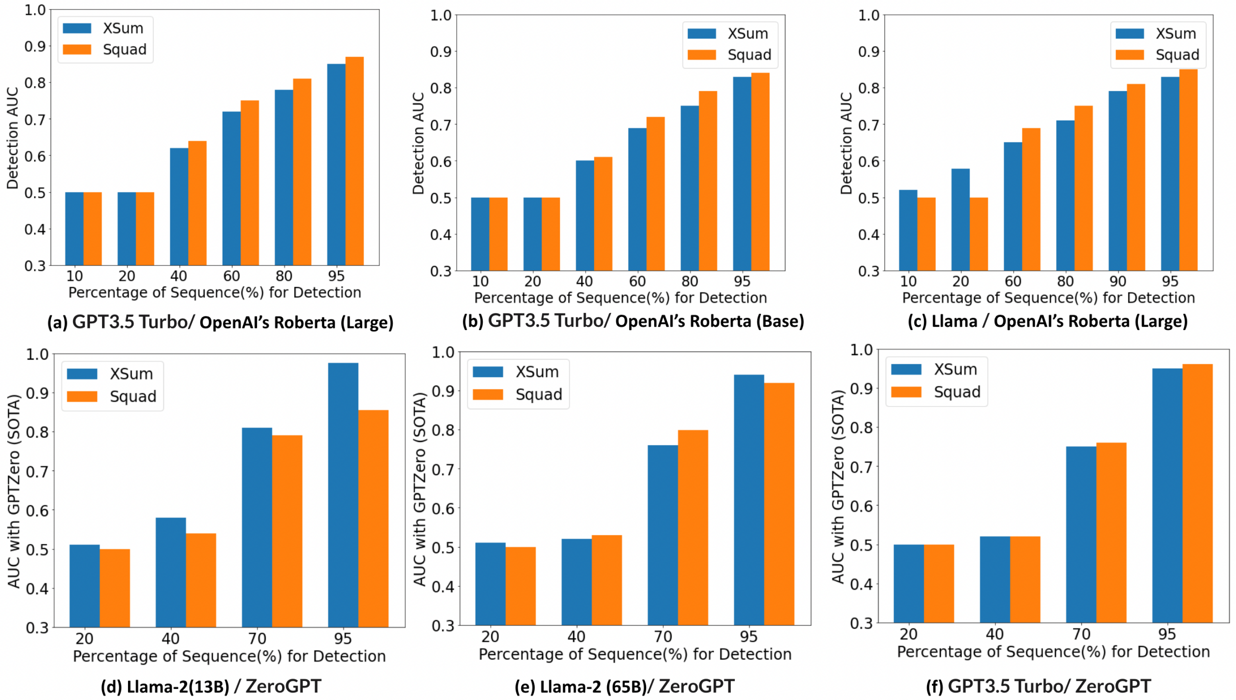

(a) GPT3.5 Turbo/ OpenAI's Roberta (Large)

(b) GPT3.5 Turbo/ OpenAI's Roberta (Base)

(c) Llama / OpenAI's Roberta (Large)

(d) Llama-2(13B) / ZeroGPT

(e) Llama-2 (65B)/ ZeroGPT

(f) GPT3.5 Turbo/ ZeroGPT

Figure 3: **(a)-(f)** validates our theorem for real human-machine classification datasets generated with XSum & Squad, with zero-shot detection performance. We use different generator/detector pairs to show the performance comparisons. For instance, (a) shows the detection performance (AUROC) of OpenAI's Roberta detector (Large) on the text generated by GPT3.5 Turbo, and we extend it to other pairs in (b)-(f). We observe that with the increase in the number of samples or sequence length for detection, the zero-shot detection performance from both the models improves from around 50% to 90% for both Xsum and Squad human-machine datasets. We also performed similar experiments with GPT-2 as well and results are available in Figure 9 in the appendix.

impressive performance aligns with our claims and provides evidence that designing a detector with high performance for AI-generated text is always possible.

(2) Detection with pairwise IID Samples: We also design an experiment where we assume that one can have access to 2 iid samples (from machine or human) for detection instead of just one example, which is practical and can be easily obtained in several scenarios. For example, consider detecting fake news or propaganda from a Twitter bot. We restructure our training set of the human-machine dataset by constructing pairwise training samples with labels of humans and machines and perform binary classification with only 30% of the enhanced pairwise dataset with very limited bag-of-word based features and Logistic regression, as shown in Figure 2c. We note that there is a statistically significant boost in detection performance with pairwise samples, even with a vanilla model and sampled dataset, which indicates that detection will be almost always possible in most scenarios where it is indeed crucial.

(3) Zero-Shot detection performance: Next, we substantiate our claims using zero-shot detection performance on the human-machine dataset for both Xsum and Squad demonstrated in Figures 3(a)-(f). For the zero-shot detection in Figures 3(a)-(c), we use the RoBERTa-Large-Detector and RoBERTa-Base-Detector from OpenAI, which are trained or fine-tuned for binary classification with datasets containing human and AI-generated texts (AIT, b). We also perform experiments with another state-of-the-art detector called ZeroGPT (AIT, a) shown in Figures 3(d)-(f). We observe that with the increase in the number of samples or sequence length of detection, the zero-shot detection performance of models improves drastically from around 50% to 90% on both Xsum and Squad human-machine datasets. Naturally, the performance of RoBERTa-Large-Detector is better compared to RoBERTa-Base-Detector, but still, the improvement in `AUROC` with the number of samples/sequence length is significant with both the models, validating our claims.

(4) Detection with Paraphrasing: We also perform the experiments with paraphrasing the document generated by the machine using a pre-trained Open-sourced Hugging-Face Paraphraser *Parrot* (Damodaran, 2021) which allows controlling the adequacy, fluency, and diversity of the generated text. We perform both supervised (Appendix), with pairwise IID Samples (Appendix) and Zero-shot detection with OpenAI's RoBERTa-Large-Detector. It is evident from Figure 4 that the detection performance decreases with paraphrasing as also shown in (Sadasivan et al., 2023; Krishna et al., 2023).

Although the detection performance drops by approximately 15% due to paraphrasing, the trend of performance improvement still remains prominent as the sequence length increases, which validates our hypothesis even under attack. Hence, one can potentially evade such attacks by considering larger sequence lengths with the sample complexity trade-off. Additionally, we observed that the performance degradation is much lesser with pairwise iid samples, highlighting the possibilities with fine-grained detectors.

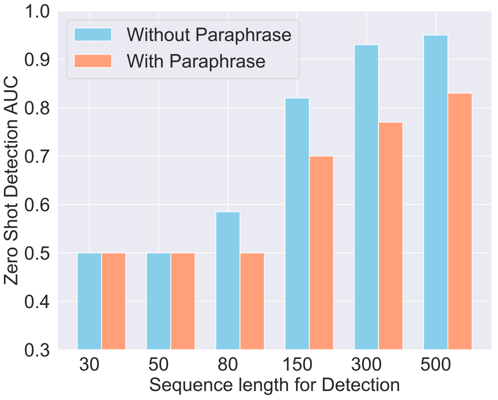

Figure 4: This figure demonstrates zero-shot detection performance with and without paraphrasing using RoBERTa-Large-Detector. Although the detection performance drops by approximately 15% due to paraphrasing, the trend of performance improvement holds as the sequence length increases.

5 Conclusion

We note that it becomes harder to detect the AI-generated text when $m(s)$ is close to $h(s)$, and paraphrasing or successive attacks can indeed reduce the detection performance as shown in our experiments. However, in several domains where we assert that by collecting more samples/sentences it will be possible to increase the attainable area under the receiver operating characteristic curve (`AUROC`) sufficiently greater than $1/2$, and hence make the detection possible. We further remark that it would be quite difficult to make LLMs exactly equal to human distributions due to the vast diversity within the human population, which may require a large number of samples from an information-theoretic perspective and provide a lower bound on the closeness distance to human distributions. While there are potential risks associated with detectors, such as misidentification and false alarms, we believe that the ideal approach is to strive for more powerful, robust, fair, and better detectors and more robust watermarking techniques. To that end, we are hopeful, based on our results, that text detection is indeed possible under most of the settings and that these detectors could help mitigate the misuse of LLMs and ensure their responsible use in society.

References

AI-based Text Detection (zerogpt). *AI Text Detector*, Jan a. URL https://www.zerogpt.com.

Ai-based text detection (openai). *AI Text Detector*, Jan b. URL https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

Scott Aaronson. My Projects at OpenAI, Nov 2022. URL https://scottaaronson.blog/?p=6823.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances, 2022.

Mikhail J. Atallah, Victor Raskin, Michael Crogan, Christian Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *Proceedings of the 4th International Workshop on Information Hiding*, IHW '01, pp. 185–199, Berlin, Heidelberg, 2001. Springer-Verlag. ISBN 3540427333.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

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. Language models are few-shot learners, 2020.