

Figure 9: **(a)-(b)** validates our theorem for real human-machine classification datasets generated with XSum & Squad, with zero-shot detection performance. We use the RoBERTa-Base-Detector (9a) and RoBERTa-Large-Detector (9b) from OpenAI which are trained or fine-tuned for binary classification with datasets containing human and AI-generated texts. 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 drastically from around 50% to 97% for both Xsum and Squad human-machine datasets.

more samples should help to perform the AI-generated text detection.

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. We believe that addressing issues such as representation space, robust watermarks, and interpretability is crucial for the safe and trustworthy application of generative language models and detection. 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.

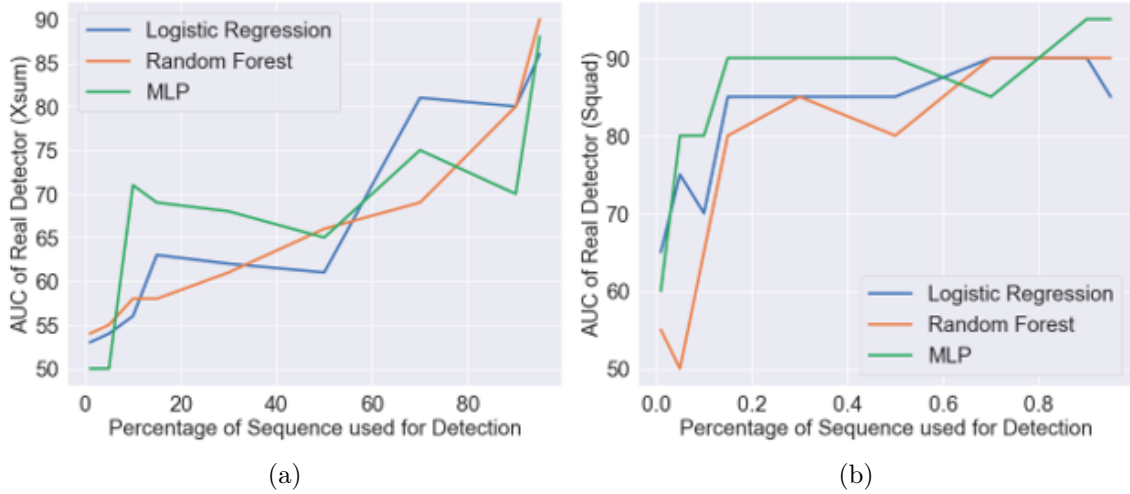


Figure 10: (a) demonstrates the detection performance of Vanilla classifiers/detectors on the Xsum dataset (Randomly sampled) generated by GPT-2. (b) demonstrates the detection performance of Vanilla classifiers/detectors on the Squad dataset (Randomly sampled) generated by GPT-2. This shows that even for vanilla detectors, our result holds for random subsets of the data.