Appendix

Table of Contents

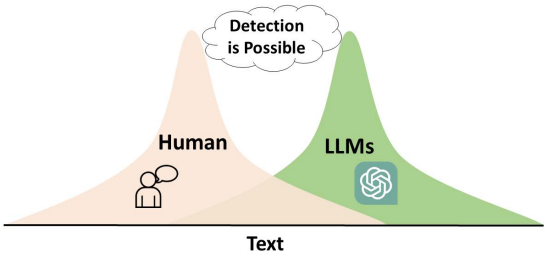
Appendix

A Additional Insights and Remarks

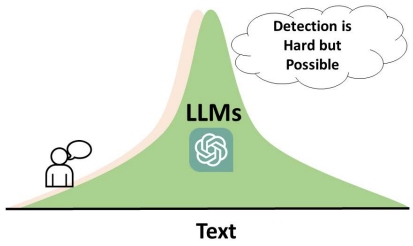
Remark 1: Insights for watermark design. From a practical perspective, even though Theorem 1 shows that detection is always possible by collecting more samples, it might be costly as well if the number n needed is extremely high. However, one could mitigate this trade-off by developing efficient watermarking techniques as discussed in (Kirchenbauer et al., 2023; Aaronson, 2022), which essentially increases the Chernoff information, or in other words, increases the δ, eventually reducing the required number of samples. Nevertheless, empirical demonstrations in (Sadasivan et al., 2023; Krishna et al., 2023) exposed the vulnerability of the watermark-based detectors with paraphrasing-based attacks, raising a genuine concern in the community about the detection of AI-generated texts.

To address this concern, more recently, interesting work by (Krishna et al., 2023) proposed a novel defense mechanism based on information retrieval principles to combat prior attacks and demonstrated its effectiveness even with a corpus size of 15M generations. This result also supports our theory, indicating that it is always possible to detect AI-generated text depending on the detection method. In addition, there are some recent open-sourced text detection tools (AIT, a,b) whose performances are also worth considering and validate the fact that detection is indeed possible under certain settings. We believe that with the new insights from this work, one can design more efficient and robust watermarks spanning a larger corpus of text, which will be hard to remove via vanilla paraphrasers.

Remark 2: Insights for detector design. This work demonstrates that detecting AI-generated text should be almost always possible but one would need to collect more samples depending on the hardness of the problem (controlled by the closeness of human and machine distributions). The recent study by (Liang et al., 2023) raises an important concern regarding the bias in some of the existing detectors. The authors in Liang et al. (2023) revealed that a significant proportion of the current detectors inaccurately classify non-native English writing samples as AI-generated, potentially leading to unjust consequences in various contexts. Interestingly, updating text generated by non-native speakers with prompts such as *Enhance it to sound more like that of a native speaker* leads to a substantial decrease in misclassification. This evidence suggests that most current detectors prioritize low perplexity as a crucial criterion for identifying a text as AI-generated, which might be flawed in various contexts, for example - academic papers as shown in (Liang et al., 2023). More specifically, we want to highlight the potential for bias in detectors relying primarily on perplexity scores, as elaborated in (Liang et al., 2023), underscoring the need for a comprehensive and equitable redesign that takes into account other relevant metrics. Our research demonstrates a promising approach to text detection, wherein the collection of more samples and the development of a multi-sample-based detector significantly enhance performance from the best word-level detector, as demonstrated by our experimental results depicted in Figures 6-8. While our results demonstrate the potential for improved detection accuracy at the paragraph level, it is important to note that this approach requires designing detectors capable of processing multiple samples. For instance, in our IMDb example, we developed a paragraph-level detector that can take the entire paragraph as input, in contrast to the word-level detector, which only processes one word at a time. Thus our approach requires the detector to deal with n samples, which may be complicated compared to processing just one sample, leading to a trade-off that could be critical for accurate detection in practice. To summarize, our work offers valuable insights into detector design, specifically about the sample complexity of AI-text detention and its connection to Chernoff information of human and machine



(a) LLM learns different distribution.



(b) LLM learns similar distribution.

Figure 5: We present the two detectability regimes for LLMs. Figure 5(a) denotes the scenario in which, when LLMs learn a different distribution, and the detection is easy. Figure 5(b) shows a scenario when LLMs' distribution is very close to human's, it is hard but possible to detect in this setting via collecting more samples. Additionally in scenarios of Figure 5(b), efficient watermarking techniques such as (Kirchenbauer et al., 2023; Krishna et al., 2023) could help in improving the separability and detectability.

distributions. We can utilize these insights to develop robust and fair detectors that enhance the overall accuracy of text detection methods.

Remark 3: Task-specific detectability & optimistic view of LLMs. In addition to our findings on the detectability of LLM-generated content, we want to highlight the significance of task-specific detectability (Figure 5). While the primary focus of Theorem 1 is to detect machine-generated text, it is important to consider the broader context of LLMs and their potential positive applications. LLMs have demonstrated significant potential to assist in a variety of tasks, including language translation (Vaswani et al., 2017), text summarization (Rush et al., 2015), dialogue systems (Serban et al., 2015), question answering (Pandya & Bhatt, 2021; Wang, 2022; Karpukhin et al., 2020), information retrieval (Chowdhury et al., 2022; Zheng et al., 2022; Kim et al., 2022), recommendation engine (Kang & McAuley, 2018; Brown et al., 2020), language grounded robotics (Ahn et al., 2022) and many others. In these scenarios, the goal is to generate high-quality text that meets the needs of the user rather than to deceive or mislead. For example, consider the application of an LLM as a tool to assist individuals or groups with moderate English writing skills to improve their writing. In this case, a well-trained LLM model could have a better (and different) distribution across $\mathcal{S}$ than the human distribution $h(s)$. This difference in distributions ensures that it should be possible to detect that AI generates the text. This understanding of detectability underscores the complexity of working with LLMs and emphasizes the importance of tailored approaches to maximize their potential. Our work provides insights into the intricacies of LLM-generated content detection, paving the way for more targeted and practical applications of these powerful models.

Remark 4: Realistic scenarios where $m(s)$ and $h(s)$ are different. Theorem 1 suggests that even small differences between the machine-generated text $m(s)$ and the human-generated text $h(s)$ should help for AI-generated text detection. In many practical applications, this difference can be easily achieved since we can control $m(s)$, but not necessarily $h(s)$. One such application is the use of LLMs to address biases and prejudices in human-generated text. While biases can arise due to the diverse backgrounds of certain communities or clusters of humans, LLMs can be trained to generate unbiased text by minimizing the likelihood of biased language in the training data. This can lead to a more inclusive and equitable society, where language use is free from discrimination.