# On the Possibilities of AI-Generated Text Detection

Souradip Chakraborty*, Amrit Singh Bedi*, Sicheng Zhu, Bang An,
Dinesh Manocha, and Furong Huang

**Abstract**

Our work addresses the critical issue of distinguishing text generated by Large Language Models (LLMs) from human-produced text, a task essential for numerous applications. Despite ongoing debate about the feasibility of such differentiation, we present evidence supporting its consistent achievability, except when human and machine text distributions are indistinguishable across their entire support. Drawing from information theory, we argue that as machine-generated text approximates human-like quality, the sample size needed for detection increases. We establish precise sample complexity bounds for detecting AI-generated text, laying groundwork for future research aimed at developing advanced, multi-sample detectors. Our empirical evaluations across multiple datasets (Xsum, Squad, IMDb, and Kaggle FakeNews) confirm the viability of enhanced detection methods. We test various state-of-the-art text generators, including GPT-2, GPT-3.5-Turbo, Llama, Llama-2-13B-Chat-HF, and Llama-2-70B-Chat-HF, against detectors, including oBERTa-Large/Base-Detector, GPTZero. Our findings align with OpenAI's empirical data related to sequence length, marking the first theoretical substantiation for these observations.

## 1 Introduction

Large Language Models (LLMs) like GPT-3 mark a significant milestone in the field of Natural Language Processing (NLP). Pre-trained on vast text corpora, these models excel in generating contextually relevant and fluent text, advancing a variety of NLP tasks including language translation, question-answering, and text classification. Notably, their capacity for zero-shot generalization obviates the need for extensive task-specific training. Recent research by (Shin et al., 2021) further highlights the LLMs' versatility in generating diverse writing styles, ranging from academic to creative, without the need for domain-specific training. This adaptability extends their applicability to various use-cases, including chatbots, virtual assistants, and automated content generation.

However, the advanced capabilities of LLMs come with ethical challenges (Bommasani et al., 2022). Their aptitude for generating coherent, contextually relevant text opens the door for misuse, such as the dissemination of fake news and misinformation. These risks erode public trust and distort societal perceptions. Additional concerns include plagiarism, intellectual property theft, and the generation of deceptive product reviews, which negatively impact both consumers and businesses. LLMs also have the potential to manipulate web content maliciously, influencing public opinion and political discourse.

*Equal Contributions. The authors are with the Department of Computer Science, University of Maryland, College Park, MD, USA. Email: {schkra3,amritbd,sczhu,bangan,dmanocha,furongh}@umd.edu

Given these ethical concerns, there is an imperative for the responsible development and deployment of LLMs. The ethical landscape associated with these models is complex and multifaceted. Addressing these challenges is vital for harnessing the societal benefits that responsibly deployed LLMs can offer. To this end, recent research has pivoted towards creating detectors capable of distinguishing text generated by machines from that authored by humans. These detectors serve as a safeguard against the potential misuse of LLMs. One central question underpinning this area of research is:

*"Is it possible to detect the AI-generated text in practice?"*

Our work provides an affirmative answer to this question. Specifically, we demonstrate that detecting AI-generated text is nearly always feasible, provided multiple samples are collected, as illustrated in Figure 1. The necessity for collecting multiple samples is consistent with real-world settings where

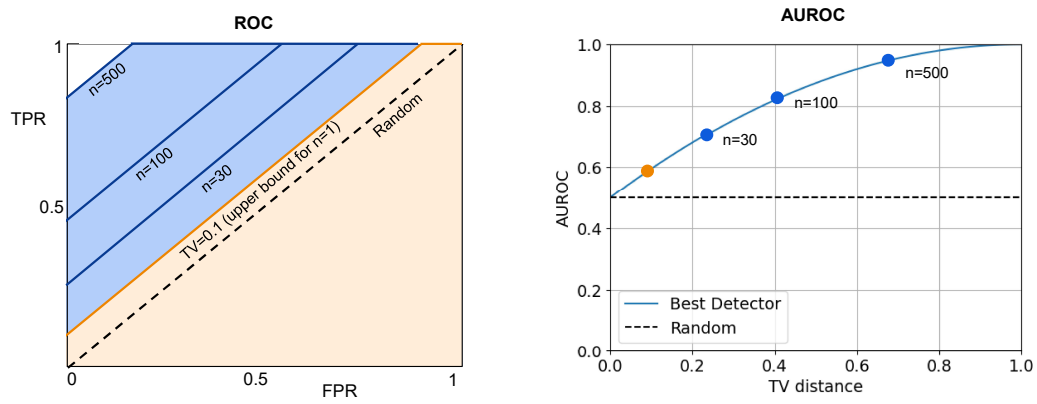


Figure 1: In light of the sample complexity bound presented in Theorem 1, we show here pictorially how increasing the number of samples $n$ used for detection would affect the ROC of the best possible detector, which is achieved by the likelihood-ratio-based classifier. We note that in the ROC curve on the left for $\mathtt{TV}(m, h) = 0.1$, the $\mathtt{AUROC}$ of the best possible detector will be 0.6 as derived in (Sadasivan et al., 2023) (shown by an orange dot in right figure). The $\mathtt{AUROC}$ of 0.6 would lead to the conclusion that detection is hard. In contrast, we note that by increasing the number of samples $n$, the ROC upper bound starts increasing towards 1 exponentially fast (shown by the shaded blue region in the left figure for different values of $n$), and hence the $\mathtt{AUROC}$ of the best possible detector also starts increasing as shown by corresponding blue dots in the right figure. This ensures that the detection should be possible even in hard scenarios when $\mathtt{TV}(m, h)$ norm is small.

data is abundant. For example, in the context of social media bot identification, one can readily collect multiple posts to determine their origin, whether machine-generated or human-authored. This real-world applicability emphasizes the importance and urgency of developing sophisticated detection mechanisms for the ethical use of LLMs. We summarize our main contributions as follows.

(1) **Possibility of AI-generated text detection**: We utilize a mathematically rigorous approach to answer the question of the possibility of AI-generated text detection. We conclude that there is a *hidden possibility* of detecting the AI text, which improves with the text sequence (token) length.

(2) **Sample complexity of AI-generated text detection**: We derive the sample complexity bounds, a first-of-its-kind tailored for detecting AI-generated text for both IID and non-IID settings.

(3) **Comprehensive Empirical Evaluations:** We have conducted extensive empirical evaluations for real datasets Xsum, Squad, IMDb, and Fake News dataset with state-of-the-art generators

(GPT-2, GPT3.5 Turbo, Llama, Llama-2 (13B), Llama-2 (70B)) and detectors (OpenAI's Roberta (large), OpenAI's Roberta (base), and ZeroGPT (SOTA Detector)).

## 2 Background on AI-Generated Text Detectors and Related Works

Recent research has shown promising results in developing detection methods. Some of these methods use statistical approaches to identify differences in the linguistic patterns of human and machine-generated text. We survey the existing approaches here.

**Traditional approaches.** They involve statistical outlier detection methods, which employ statistical metrics such as entropy, perplexity, and $n$-gram frequency to differentiate between human and machine-generated texts (Lavergne et al., 2008; Gehrmann et al., 2019). However, with the advent of ChatGPT (OpenAI) (OpenAI, 2023), a new innovative statistical detection methodology, DetectGPT (Mitchell et al., 2023), has been developed. It operates on the principle that text generated by the model tends to lie in the negative curvature areas of the model's log probability. DetectGPT (Mitchell et al., 2023) generates and compares multiple perturbations of model-generated text to determine whether the text is machine-generated or not based on the log probability of the original text and the perturbed versions. DetectGPT significantly outperforms the majority of the existing zero-shot methods for model sample detection with very high `AUC` scores (note that we use the terms `AUROC` and `AUC` interchangeably for presentation convenience).

**Classifier-based detectors.** In contrast to statistical methods, classifier-based detectors are common in natural language detection paradigms, particularly in fake news and misinformation detection (Schildhauer, 2022; Zou & Ling, 2021). OpenAI has recently fine-tuned a GPT model (OpenAI, 2023) using data from Wikipedia, WebText, and internal human demonstration data to create a web interface for a discrimination task using text generated by 34 language models. This approach combines a classifier-based approach with a human evaluation component to determine whether a given text was machine-generated or not. These recent advancements in the field of detecting AI-generated text have significant implications for detecting and preventing the spread of misinformation and fake news, thereby contributing to the betterment of society (Schildhauer, 2022; Zou & Ling, 2021; Kshetri & Voas, 2022).

**Watermark-based identification.** An alternative detection paradigm that has garnered significant interest in this field is the evolution of watermark-based identification (Verma et al., 2009; Wadhera et al., 2022). One of the most exciting works in recent times around this research revolves around watermarking and developing efficient watermarks for machine-generated text detection. Historically, watermarks have been employed in the realm of image processing and computer vision to safeguard copyrighted content and prevent intellectual property theft (Langelaar et al., 2000). They can also be used for data hiding, where information is hidden within the watermark itself, allowing for secure and discreet transmission of information. Early research by (Atallah et al., 2001; Meral et al., 2009) was among the first to demonstrate the potential of watermarks in language through syntax tree manipulations. More recently with the advent of ChatGPT, innovative work by (Kirchenbauer et al., 2023) has shown how to incorporate watermarks by using only the LLM's logits at each step. The watermarking technique proposed by (Kirchenbauer et al., 2023) allows for the verification of a watermark's authenticity by employing a specific hash function. More specifically, the soft watermarking approach by (Kirchenbauer et al., 2023) involves categorizing tokens into "green" and "red" lists for generating distinct patterns. Watermarked language models are more likely to