select tokens from the green list, based on prior tokens, resulting in watermarks that are typically unnoticeable to humans. These advancements in watermarking technology not only strengthen copyright protection and content authentication but also open up new avenues for research in areas such as privacy in language, secure communication, and digital rights management.

**Impossibility result**. The interesting recent literature by (Sadasivan et al., 2023; Krishna et al., 2023) showed the vulnerabilities of watermark-based detection methodologies using vanilla paraphrasing attacks. (Sadasivan et al., 2023) developed a lightweight neural network-based paraphraser and applied it to the output text of the AI-generative model to evade a whole range of detectors, including watermarking schemes, neural network-based detectors, and zero-shot classifiers. (Sadasivan et al., 2023) also introduced a notion of spoofing attacks where they exposed the vulnerability of LLMs protected by watermarking under such attacks. (Krishna et al., 2023) on the other hand, trained a paraphrase generation model capable of paraphrasing paragraphs and showed that paraphrased texts with DIPPER (Krishna et al., 2023) evade several detectors, including watermarking, GPTZero, DetectGPT, and OpenAI's text classifier with a significant drop in accuracy. Additionally, (Sadasivan et al., 2023) highlighted the impossibility of machine-generated text detection when the total variation (TV) norm between human and machine-generated text distributions is small.

In this work, we show that there is a ***hidden possibility*** of detecting the AI-generated text even if the TV norm between human and machine-generated text distributions is small. This result is in support of the recent detection possibility claims by Krishna et al. (2023).

# 3 Proposed Approach: Methodology and Analysis

## 3.1 Notations and Definitions

Before discussing the main results, let us define the notations used in this paper. We define the set of all possible texts (textual representations) as $\mathcal{S}$, a human-generated text distribution as $h(s)$ over $s \in \mathcal{S}$, and machine-generated text distribution as $m(s)$. Here $m(s)$ and $h(s)$ are valid probability density functions. We can also modify the same notation given a specific prompt (noted by $p$) or context (denoted by $c$) or question (denoted by $q$) accordingly, such as $\mathcal{S}_c$, $h(s \mid p, c, q)$, and $m(s \mid p, c, q)$ respectively. However, for the sake of clarity and ease of discussion in this work, we will omit the use of complex notation.

In the literature, the problem of detecting AI-generated text is considered as a binary classification problem. The (potentially nonlinear and complex) detector $D(s)$ maps the sample $s \in \mathcal{S}$ to $\mathbb{R}$ for possible binary classification, and then compares it against a threshold $\gamma$ to perform detection. $D(s) \geq \gamma$ is classified as AI-generated while $D(s) < \gamma$ is categorized as human-generated. For the detector $D(s)$ to detect whether the text samples $s$ is generated from the machine or not, we need to study the receiver operating characteristic curve (ROC curve) (Fawcett, 2006), which involves two terms, namely True Positive Rate (TPR) and False Positive Rate (FPR). Once we obtain ROC, we can study the area under the ROC curve AUROC, which characterizes the detection performance of detector $D$. The upper bound on AUROC describes the performance of the best possible detector.

Under a detection threshold $\gamma$, TPR and FPR are denoted as $\text{TPR}_\gamma$ and $\text{FPR}_\gamma$ respectively:

$$\text{TPR}_\gamma \text{ :Probability of detecting AI-generated text as AI-generated under threshold } \gamma, \tag{1}$$

$$\text{FPR}_\gamma \text{ :Probability of detecting human-generated text as AI-generated under threshold } \gamma. \tag{2}$$

The rigorous definitions of $\mathrm{TPR}_\gamma$ and $\mathrm{FPR}_\gamma$ are as follows.

$$\mathrm{TPR}_\gamma = \mathbb{P}_{s\sim m(\cdot)}[D(s)\geq\gamma] = \int \mathbb{I}_{\{D(s)\geq\gamma\}}\cdot m(s)\cdot ds, \tag{3}$$

$$\mathrm{FPR}_\gamma = \mathbb{P}_{s\sim h(\cdot)}[D(s)\geq\gamma] = \int \mathbb{I}_{\{D(s)\geq\gamma\}}\cdot h(s)\cdot ds, \tag{4}$$

where $\mathbb{I}_{\{\text{condition}\}}$ is the indicator function which takes value 1 if the condition is true, and 0 otherwise. Note that without loss of generality, we have chosen to consider $m(s)$ and $h(s)$ as the probability density function of machine and human on a sample $s$ by considering continuous $s$ (as also considered in (Sadasivan et al., 2023), but similar results hold for discrete $s$ by replacing the integral with a summation and by considering $m(s)$ and $h(s)$ as the probability mass function of machine and human on a sample $s$.

Both $\mathrm{TPR}_\gamma$ and $\mathrm{FPR}_\gamma$ are within the closed interval [0, 1] for any threshold $\gamma$. For a good detector, $\mathrm{TPR}_\gamma$ should be as high as possible, and $\mathrm{FPR}_\gamma$ should be as low as possible. As a result, a high *area under the ROC curve* (AUROC) is desirable for detection. AUROC is between $1/2$ and 1, i.e., AUROC $\in [1/2, 1]$. An AUROC value of $1/2$ means a random detection and a value of 1 indicates a perfect detection. For efficient detection, the goal is to design a detector $D$ such that AUROC is as high as possible.

## 3.2 Hidden Possibilities of AI-Generated Text Detection

To study the AUROC for any detector $D$, we start by invoking LeCam's lemma (Le Cam, 2012; Wasserman, 2013) which states that for any distributions $m$ and $h$, given an observation $s$, the minimum sum of Type-I and Type-II error probabilities in testing whether $s\sim m$ versus $s\sim h$ is equal to $1-\mathsf{TV}(m,h)$. Hence, mathematically, we can write

$$\underbrace{\mathbb{P}_{s\sim h(\cdot)}[D(s)\geq\gamma]}_{\text{Type-I error (false positive)}} + \underbrace{\mathbb{P}_{s\sim m(\cdot)}[D(s)<\gamma]}_{\text{Type-II error (false negative)}} \geq 1-\mathsf{TV}(m,h), \tag{5}$$

for any detector $D$ and any threshold $\gamma$. We note that the above bound is tight and can always be achieved with equality by likelihood-ratio-based detectors for *any* distribution $m$ and $h$, by the Neyman-Pearson Lemma (Cover, 1999, Chapter 11). We restate the lemma for completeness and discuss its tightness in Appendix B.1. From the definitions of TPR and FPR in (3)-(4), it holds that

$$\mathrm{FPR}_\gamma + 1 - \mathrm{TPR}_\gamma \geq 1-\mathsf{TV}(m,h), \tag{6}$$

which implies that

$$\mathrm{TPR}_\gamma \leq \min\{\mathrm{FPR}_\gamma + \mathsf{TV}(m,h), 1\}, \tag{7}$$

where min is used because $\mathrm{TPR}_\gamma \in [0,1]$. The upper bound in (7) is called the ROC upper bound and is the bound leveraged in one of the recent works (Sadasivan et al., 2023) to derive AUROC upper bound $\mathsf{AUC} \leq \frac{1}{2} + \mathsf{TV}(m,h) - \frac{\mathsf{TV}(m,h)^2}{2}$ which holds for any $D$. This upper bound led to the claim of the impossibility of detecting the AI-generated text whenever $\mathsf{TV}(m,h)$ is small.

**Hidden Possibility.** However, we note that the claim of impossibility from the AUROC upper bound could be too conservative for detection in practical scenarios. For instance, we provide a ***motivating example*** of detecting whether an account on Twitter is an AI-bot or human. It is natural that we will have a collection of text samples from the account, denoted by $\{s_i\}_{i=1}^n$, and it is realistic to assume that $n$ is very high. Therefore, the natural practical question is whether we can detect if the provided text set $\{s_i\}_{i=1}^n$ is machine-generated or human-generated. With this motivation, we next explain that detection is always possible.

We formalize the problem setting and prove our claim by utilizing the existing results in the information theory literature. Let us consider the same setup as detailed before, while we are given a set of samples $S := \{s_i\}_{i=1}^n$. For simplicity, we assume that the samples are i.i.d. drawn from either the human $h$ or machine $m$. Interestingly, now the hypothesis test can be re-written as

$$H_0 : S \sim m^{\otimes n} \quad \texttt{v.s.} \quad H_1 : S \sim h^{\otimes n}, \tag{8}$$

where $m^{\otimes n} := m \otimes m \otimes \cdots \otimes m$ ($n$ times) denotes the product distribution, as does $h^{\otimes n}$. This is one of the key observations that focus on the correct hypothesis-testing framework with multiple samples. Similar to before (cf. 7), based on Le Cam's lemma, it holds that now $1 - \mathsf{TV}(m^{\otimes n}, h^{\otimes n})$ gives the minimum Type-I and Type-II error rate, which implies

$$\mathrm{TPR}_\gamma^n \leq \min\{\mathrm{FPR}_\gamma^n + \mathsf{TV}(m^{\otimes n}, h^{\otimes n}), 1\}, \tag{9}$$

where

$$\mathrm{TPR}_\gamma^n = \mathbb{P}_{S\sim m^{\otimes n}}[D(S) \geq \gamma] = \int \mathbb{I}_{\{D(S)\geq\gamma\}} \cdot m^{\otimes n}(S) \cdot dS, \tag{10}$$

$$\mathrm{FPR}_\gamma^n = \mathbb{P}_{S\sim h^{\otimes n}}[D(S) \geq \gamma] = \int \mathbb{I}_{\{D(S)\geq\gamma\}} \cdot h^{\otimes n}(S) \cdot dS. \tag{11}$$

We emphasize that the term $\mathsf{TV}(m^{\otimes n}, h^{\otimes n})$ is an increasing sequence in $n$ and eventually converges to 1 as $n \to \infty$. Due to the data processing inequality, it holds that $\mathsf{TV}(m^{\otimes k}, h^{\otimes k}) \leq \mathsf{TV}(m^{\otimes n}, h^{\otimes n})$ when $k \leq n$ and naturally leads to $\mathsf{TV}(m, h) \leq \mathsf{TV}(m^{\otimes n}, h^{\otimes n})$. This is a crucial observation, showing that even if the machine and human distributions were close in the sentence space, by collecting more sentences, it is possible to inflate the total variation norm to make the detection possible.

Now, from the large deviation theory, we can show that the rate at which total variation distance approaches 1 is exponential with the number of samples (Polyanskiy & Wu, 2022, Chapter 7),

$$\mathsf{TV}(m^{\otimes n}, h^{\otimes n}) = 1 - \exp\left(-nI_c(m, h) + o(n)\right), \tag{12}$$

where, $I_c(m, h)$ is known as the *Chernoff information* and is given by $I_c(m, h) = -\log \inf_{0\leq\alpha\leq1} \int m^\alpha(s)h^{1-\alpha}(s)ds$. The above expressions lead to Proposition 1 next.

**Proposition 1** (**Area Under ROC Curve**). *For any detector $D$, with a given collection of i.i.d. samples $S := \{s_i\}_{i=1}^n$ either from human $h(s)$ or machine $m(s)$, it holds that*

$$\mathsf{AUROC} \leq \frac{1}{2} + \mathsf{TV}(m^{\otimes n}, h^{\otimes n}) - \frac{\mathsf{TV}(m^{\otimes n}, h^{\otimes n})^2}{2}, \tag{13}$$

*where $\mathsf{TV}(m^{\otimes n}, h^{\otimes n}) := 1 - \exp\left(-nI_c(m, h) + o(n)\right)$ and $I_c(m, h)$ is the Chernoff information. Therefore, the upper bound of $\mathsf{AUROC}$ increases exponentially with respect to the number of samples $n$.*

The proof of the above proposition follows by integrating the $\mathrm{TPR}_\gamma^n$ upper bound in (9) over $\mathrm{FPR}_\gamma^n$. We note that the expression in (13) and the equality of $\mathsf{TV}$ distance in terms of Chernoff information presents an interesting connection between the number of samples and $\mathsf{AUROC}$ of the best possible detector (which archives the bound in (9) with equality). It is evident that if we increase the number of samples, $n \to \infty$, the total variation distance $\mathsf{TV}(m^{\otimes n}, h^{\otimes n})$ approaches 1 and that too exponentially fast, and hence increasing the $\mathsf{AUROC}$. This indicates that as long as the two distributions are not exactly the same, which is rarely the same, the detection will always be possible by collecting more samples as established next.