3.3 Attainability of the AUROC Upper-Bound via Likelihood-Ratio-Based Detectors

Likelihood-ratio-based Detector. Here, we discuss the attainability of bounds in Proposition 1 to establish that the bound is indeed tight. We note that it is a well-established fact in the literature that a likelihood-ratio-based detector would attain the bound for any distributions h and m and hence is the best possible detector (detailed proof provided in Appendix B.1). We discuss the likelihood-ratio-based detector here for completeness in the context of LLMs as follows. Specifically, the likelihood ratio-based detector is given by

$$D^*(S) := \begin{cases} \text{Text from machine} & \text{if } m^{\otimes n}(S) \geq h^{\otimes n}(S), \\ \text{Text from human} & \text{if } m^{\otimes n}(S) < h^{\otimes n}(S). \end{cases} \tag{14}$$

We proved in Appendix B.1 that the detector in (14) attains the bound and is the best possible detector.

Sample Complexity of Best Possible Detector. To further emphasize the dependence on the number of samples n, we derive the sample complexity bound of AI-generated text detection in Theorem 1 as follows.

Theorem 1 (**Sample Complexity of AI-generated Text Detection (Possibility Result)**). *If human and machine distributions are close* $\mathsf{TV}(m, h) = \delta > 0$*, then to achieve an* AUROC *of* ϵ*, we require*

$$n = \Omega\left(\frac{1}{\delta^2}\log\left(\frac{1}{1-\epsilon}\right)\right) \tag{15}$$

number of samples for the best possible detector which is likelihood-ratio-based as mentioned in (14)*, for any* $\epsilon \in [0.5, 1)$*. Therefore, AI-generated text detection is possible for any* $\delta > 0$*.*

The proof of Theorem 1 is provided in Appendix B.2. From the statement of Theorem 1, it is clear that, as long as $\delta > 0$ (which means no matter how close human $h(s)$ and $m(s)$ distributions are) and $\epsilon < 1$, there exists n such that we can achieve high AUROC and perform the detection. Here, n corresponds to the number of sentences generated by either humans or machines which we need to detect. We provide additional detailed remarks and insights in Appendix A.

3.4 Extension to Non-IID case

We extend the sample complexity results of Theorem 1 to the non-iid setting in this subsection. In order to accomplish that, we make certain assumptions about the structures present in the input, which is a well-founded assumption that proves to be practical and applicable in the context of various natural language tasks (for ex: present of topics in documents (Jelodar et al., 2018; Loureiro et al., 2023)). Let us denote the strength of the association with ρ to characterize the dependence between the sequences s_i and the dependence is given as

$$\mathbb{E}[S_i | S_{i-1} = s_{i-1}, \cdots, S_1 = s_1] = \rho \frac{\sum_{k=1}^{i-1} s_k}{i-1} + (1-\rho)\mathbb{E}[S_i], \tag{16}$$

which boils down to the iid case for $\rho = 0$. An increasing ρ indicates increasing dependence on the previous sequence with $\rho = 1$ means that the conditional expectation can be completely expressed in terms of the previous samples in the sequence. The dependence assumption of (16) embodies a

natural intuition for the domain of natural language and serves as a foundation for extending our results to non-iid scenarios. Eq. (16) provides a way to measure the dependence between random variables, which is later used to extend Chernoff bound to non-iid cases. . In the context of LLMs, one can think of the sum $\sum s_k$ as the "average meaning" of these text samples, such as "woman" + "royalty" may have a similar meaning as "queen".

Before introducing the final result, let us assume the number of sequences or samples is denoted by n, there are L independent subsets, and the corresponding subset is represented by τ_j where $j \in (1, 2 \cdots, L)$, where τ_j consists of c_j samples (dependent). This is a natural assumption in NLP where a large paragraph often consists of multiple topics, and sentences for each topic are dependent. With the above definitions, we state the main result in Theorem 2 for the non-iid setting.

Theorem 2 (**Sample Complexity of AI-generated Text Detection (non-iid)**). *If human and machine distributions are close* $\mathsf{TV}(m, h) = \delta > 0$*, then to achieve an* AUROC *of* ϵ*, we require*

$$n = \Omega\left(\frac{1}{\delta^2}\log\left(\frac{1}{1-\epsilon}\right) + \frac{1}{\delta}\sum_{j=1}^{L}(c_j - 1)\rho_j + \sqrt{\frac{1}{\delta^2}\log\left(\frac{1}{1-\epsilon}\right)\cdot\frac{1}{\delta}\left(\sum_{j=1}^{L}(c_j-1)\rho_j\right)}\right) \tag{17}$$

number of samples for the best possible detector which is likelihood-ratio-based as mentioned in (14)*, for any* $\epsilon \in [0.5, 1)$*. Therefore, AI-generated text detection is possible for any* $\delta > 0$*.*

The proof is provided in Appendix 2. From the statement of Theorem 2, we note that for $\delta > 0$ ($h(s)$ and $m(s)$ are close but not exactly the same) and $\epsilon < 1$, there exists n such that we can achieve high AUROC and perform the detection. In comparison to the iid result in Theorem 1, the non-iid result in Theorem 2 has an additional term that depends on c_j and ρ_j. Clearly, for $\rho_j = 0$, the sample complexity result in Theorem 2 boils down to the result in Theorem 1.

4 Experimental Studies

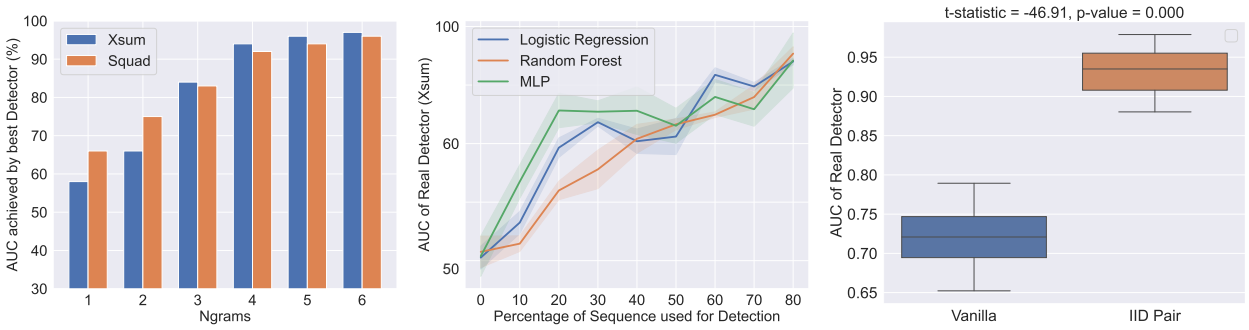
In this section, we provide detailed empirical evidence to support our detectability claims of this work. We consider various human-machine generated datasets and general language classification datasets.

AUROC Discussion and Comparisons: We first try to explain the meaning of the mathematical results we obtain via simulations. For instance, we show a pictorial representation of AUROC bound we obtained in Proposition 1 and compare it against the ROC upper bound we mentioned in (9) for different values of n. In Figure 1, we show that even if the original distributions of human $h(s)$ and machines $m(s)$ are close in TV norm $\mathsf{TV} = 0.1$, we can increase the ROC area (and hence AUROC) via increasing the number of samples we collect n to perform the detection.

4.1 Real Data Experiments

In this section, we perform a detailed experimental analysis and ablation to validate our theorem with several real human-machine generated datasets as well as general natural language datasets.

Datasets, AI-Text Generators and Detectors Description: Our experimental analysis spans across 4 critical datasets, including the news articles from XSum dataset (Narayan et al., 2018), Wikipedia paragraphs from Squad dataset (Rajpurkar et al., 2016), IMDb reviews (Maas et al., 2011), and Kaggle FakeNews dataset (Lifferth, 2018), utilizing the datasets in a diverse manner to validate our hypothesis. The first two datasets (XSum and Squad) have been leveraged to



(a) (b) (c)

Figure 2: **(a)-(c)** validates our theorem for real human-machine classification datasets generated with XSum (Narayan et al., 2018) and Squad (Rajpurkar et al., 2016), showing that with an increase in the number of samples/sequence length, detection performance improves significantly. Figure 2a shows that the AUROC achieved by the best possible detector using the equation increases significantly from 58% to 97% with an increase in the Ngrams of the feature space for both Xsum and Squad datasets. Figure 2b demonstrates the improvement in AUROC with respect to sequence length using various real detectors/classifiers. Figure 2c shows using a box-plot-based comparison that if we consider 2 iid sequences (from either machine/human) to detect instead of one, the AUROC of the real detector improves drastically from 73% to 97%, hence validating our hypothesis.

generate machine-generated text by prompting an LLM with the first 50 tokens of each article in the dataset, sampling from the conditional distribution of the LLMs, as followed in (Mitchell et al., 2023; Krishna et al., 2023; Sadasivan et al., 2023). Specifically, we use a diverse set of SOTA open-source text generators including GPT-2, GPT-3.5-Turbo, Llama, Llama-2-13B-Chat-HF, and Llama-2-70B-Chat-HF as the LLM for generating the machine-generated text using the token prompts as described above. We consider 500 passages from both the Xsum and Squad datasets and subsequently 500 machine-generated texts corresponding to them using GPT-2 and evaluate the detection performance in 3 broad categories including *(1) supervised detection, (2) contrastive with i.i.d. samples, and (3) zero-shot performance*. Finally, we leverage two additional general language datasets (detailed in Appendix C.1), IMDb and Kaggle FakeNews, to give more insights into the separability and detection performance with an increasing number of samples.

(1) Supervised detection performance: To validate our hypothesis from a supervised detection/classification perspective, we first compute the total variation distance between the human and machine-generated texts at various n-gram levels where $\texttt{n-gram} = 1$ indicates the detection is at a word level, and as we increase it, it approaches sentence to paragraph level. We subsequently estimate the AUROC of the best detector using equation (14) by increasing the length of the n-gram from 1 to 6 as shown in Figure 2a. It is evident that with increasing n-grams, the AUROC of the best detector increases significantly from 58% to 97% for both Xsum and Squad datasets. This empirical observation completely aligns with our theory and intuition. To further test our hypothesis with real detectors, we train 3 vanilla classification models including Logistic Regression, Random Forest, and a 2-layer Neural Network with TF-IDF-based feature representation (bag of words) on the human-machine generated datasets including Xsum and Squad. We report the performance of the test AUROC with increasing sequence length in Figure 2b, which shows a significant increase in accuracy as the sequence length increases even with real detectors. This observation is also supported by the results obtained from Open-AI and summarized in the report (Solaiman et al., 2019). This