MMIDR:Multi-scale Mutual Information for AI Detection via Rewriting

ZheKai Shen[1], ShiYu Chu[2]

Province/State: Shanghai, China

Advisor: Haobin Zhu

September 15, 2024

MMIDR:Multi-scale Mutual Information for AI Detection via Rewriting

ZheKai Shen[1], ShiYu Chu[2]

Abstract

Explosive hardware breakthroughs in recent years have enabled scaling laws for large models, and emergent roles have become more common, the paradigm of human-computer interaction is undergoing unprecedented changes. LLMs have shown great potential in promoting programming efficiency, optimizing text creation, and advancing scientific research processes.But if it is allowed to grow wild and unchecked, it can lead to irreversible situations. For example, the dissemination of false and low-quality information, and misleading and biased ideologies can be a recipe for disaster. Considering these concerns,the development of a dependable content censorship mechanism for information security seems to be a matter of urgency.Galvanized by this conundrum, I embarked upon a odyssey of intellectual excavation, unearthing a paradigm-shifting revelation: the unfettered ideation of the sapient mind, coupled with the highly variegated spectrum of literary acumen exhibited by our species, engenders a corpus of work that diverges from machine-generated prose in a manner so profound as to defy facile quantification. This disparity manifests with particular salience in the realms of morphological synthesis and syntactic orchestration. The textual artifacts born of human cognition evince a degree of heterogeneity that far surpasses the output of even the most sophisticated linguistic algorithms, In the pursuit of linguistic exegesis, one discerns a salient phenomenon, particularly when delving into the abstruse intricacies of morphemic amalgamation and the byzantine configuration of syntactic constituents. This observable peculiarity becomes increasingly conspicuous upon meticulous scrutiny of the paradigmatic approaches employed in lexical entity genesis and the combinatorial potentialities latent within propositional frameworks. The fundamental axiom underpinning this perceived duality may be ascribed to the ineluctable proclivity of Expansive Textual Synthesis Systems to preserve an inviolable constancy in their architectural cohesion, concomitant with an unwavering adherence to stylistic uniformity throughout the entirety of their generative machinations. This stands in stark contrast to the mercurial nature of human expression, which is subject to the vagaries of individual idiosyncrasy and the capricious influence of myriad exogenous factors., i.e., the model perceives the best practice. This best practice is favoured by the model during the rewriting process.Predicated upon the aforementioned empirical observations and theoretical exegesis, this scholarly exposition proposes to elucidate a groundbreaking methodological paradigm for detection and analysis: Multi scale Mutual Information for AI Detection via Rewriting (MMIDR). This method achieves effective recognition of AI generated content by performing the multi-scale rewriting tasks on the test texts and analyzing the information changes of samples before and after rewriting.The MMIDR approach is unique relative to some past means in that it does not introduce additional training overheads and costs., but fully utilizes the black box characteristics of LLM. By analysing the generative patterns and rules of the LLM itself, the intrinsic patterns of the text to be tested are analysed at multiple scales.

1 Introduction

Unlike discriminative AI, the power of generative AI seems to be more intimidating. A series of serious risks and challenges have also emerged.Abdali et al. (2024) In the field of cybersecurity, LLMs can be used in phishing attacks, generating misleading information and adding to the already precarious state of cybersecurity.Roy et al. (2023) A study suggests that LLM systems may face specific security threats, including prompt injection attacks, generation of harmful content, and indirect leakage of sensitive information.Cui et al. (2024); Liu et al. (2023) llm's vast store of knowledge puts many academics to shame and is a natural teaching tool, but again, it can lead to academic misconduct, as well as foster bias, provide misleading information.The illusion of LLMs cannot be effectively mitigated, and the high quality of knowledge is difficult to ensure.Mitchell et al. (2023) In addition, the use of LLMs in scientific research has also raised many issues. A study explores the risks of using LLM agents for scientific purposes, such as outdated knowledge and potential resource waste.Mozes et al. (2023); Wang et al. (2024) In healthcare, the use of LLM is even more of a thin ice to walk on, not only in terms of patient data privacy, but also in terms of healthcare's extreme need for safety, where LLM can't vouch for any of its treatment protocols, and where an incorrect output could result in the loss of a life.Tang et al. (2024) This highlights the importance of ensuring responsible deployment of AI.A paper published in Nature discusses the risks of using proprietary LLMs in research environments,more typically the lack of rigorous proof, and the bias in the scientific research process that can result from reliance on these models.Pressman et al. (2024) To address these challenges, ensuring responsible use of generative AI tools has become crucial. A paper highlights the various risks related to LLMs, including privacy issues, copyright infringement, addressing bias and misinformation,and exploring strategies to reduce these risks,such as bias correction techniques and validation measures.Van Dis et al. (2023) Recent research has focused on identifying text generated by AI.Several studies have considered the problem as a discriminative task. and constructed discriminators through deep neural networks.Das et al. (2024) However, among the latest LLMs,traditional heuristic-based detection methods are starting to become obsolete.Jawahar et al. (2020) The current technology regarding the numerical output metrics in the study of Gehrmann et al. (2019). But trying to implement similar operations in closed-source black-box models like GPT-3.5 and GPT-4 is unlikely. Therefore, it is crucial to quickly establish a set of detection methods matching the existing language modelling environment. It not only helps reduce the potential negative impact of LLMs but also facilitates the implementation of superior and advanced detection algorithms.

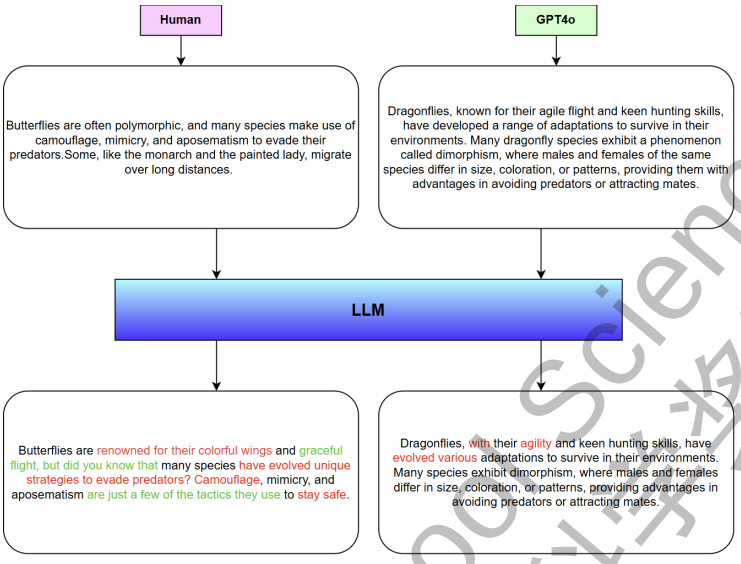

Figure 1: MMIDR Schema

2 Related Work

Since ancient times, the topic of spear and shield has always been a fascinating one, and this also applies to the field of LLM, where the demand for AIGC testing has gradually climbed due to a strong development of LLM. Researchers have all given different solutions to this problem. From supervised learning to unsupervised learning. And even statistical feature-based methods and deep learning models. Ippolito et al. (2019) conducted a comparative analysis on the capabilities of detectors and human assessors in determining which has been generated by a machine. They employed a finetuned model from BERT as the primary classifier and co-locating it with other baselines. The study revealed that automatic detectors generally outperformed human evaluators, particularly as text length increased. They also observed that different decoding strategies. For example,top-k and nucleus sampling, significantly influenced detection difficulty. This research underscored the importance of utilizing both human and automatic detectors to assess the "human-likeness" of generated text. Gallé et al. (2021) introduced a novel pervised and distributed method for detecting text generated by machines. Their approach, based on the concept of super-maximal repeats, employed multiple iterations of pseudo-labeling and classifier training. In a semi-supervised setting, the method achieved over 0.9 precision accuracy for text generated using top-k sampling. Notably, the method maintained high accuracy even in a completely unsupervised scenario. This study demonstrated that while modern language models closely mimic human text in word-level statistical features, detectable differences persist in higher-order n-gram repetition patterns. Wang et al. (2023) proposed SeqXGPT, a sentence-level AI-generated

text detection method. SeqXGPT leverages log probabilities from multiple open-source language models as features, utilizing efficient networks that integrate CNN and and self-attention to process them. In multi-model multi-classification detection tasks, SeqXGPT attained a macro F1 score of 0.957, significantly surpassing existing baseline methods.More importantly, SeqXGPT exhibited robust generalization capabilities on out-of-distribution (OOD) datasets, achieving a macro F1 score of 0.928. This research provides effective solutions for AI-generated text detection at different scales. Including sentence level and document level. Gaggar et al. (2023) compared the performance of SVM and two scales of RoBERTa models in detecting ChatGPT (GPT-3.5 turbo) generated text. Upon meticulous scrutiny across a diverse spectrum of sentential magnitudes, the empirical data evinced a hierarchical gradient of efficacy among the computational paradigms under examination. The avant-garde RoBERTa architecture exhibited superlative prowess in its operational capacities, eclipsing its progenitor, the RoBERTa-base framework, which in turn demonstrated superior performance vis-à-vis the more conventional Support Vector Machine methodology. This multifaceted analytical endeavor transcends mere juxtaposition of algorithmic competencies; it delves into the nuanced interplay between textual prolixity and detection acuity, thereby furnishing a rich tapestry of heuristic insights poised to inform and guide subsequent scholarly pursuits in this burgeoning field of inquiry. Mitchell et al. (2023) introduced DetectGPT, a zero-shot detection method based on the local curvature of language model log probability functions. By calculating the log probability differences between original texts and the scrambled version. DetectGPT effectively distinguishes between AI-generated and human-generated text. Across multiple datasets and models, DetectGPT exceeded the performance of current zero-shot detection approaches,and in certain cases it surpassed them. The main benefit of this approach lies in its versatility and ability to be adapted to various decoding scenarios. Shah et al. (2023) employed multiple ML classification algorithms in conjunction with explainable AI (xAI) techniques to detect AI-generated text. By utilizing various textual features,they were able to achieve a classification accuracy of 0.93 on the task. The study also utilized LIME and SHAP techniques, in order to clarify the model's decision-making process, key features such as Herdan's C, MaaS, and Simpson's Index will be identified. The research led to advancements in detection accuracy as well as enhanced model interpretability.

3 Method

In this section, we present the Multi-scale Mutual Information for AI Detection via Rewriting (MMIDR) method for distinguishing between human-authored and AI-generated text. We first present the foundations and key definitions, followed by the detailed formulation of MMIDR. We then examine the theoretical assurances and address the practical execution of the technique.

3.1 Definitions

Let $(\Omega, \mathcal{F}, P)$ be a probability space. X be a random variable on this space taking values in a measurable space (S, Σ). We begin by introducing several key concepts:
Let $\{\mathcal{F}_\tau\}_{\tau \geq 0}$ be a filtration, i.e., an increasing family of σ-algebras generated by the first τ tokens of the context.

Kullback-Leibler Divergence: For probability measures P and Q on $(\Omega, \mathcal{F})$, K-L divergence is defined as:

$$D_{KL}(P\|Q) = \int_\Omega \log\left(\frac{dP}{dQ}\right) dP$$

where $\frac{dP}{dQ}$ is the Radon-Nikodym derivative.

Mutual Information: The mutual information between X and Y is defined as:

$$I(X;Y) = D_{KL}(P_{XY}\|P_X \otimes P_Y)$$

Rewriting Operator: Let R be a rewriting operator that maps a text X to its rewritten version $R(X)$.

3.2 MMIDR Formulation

We then present the definition of MMIDR and its key properties.

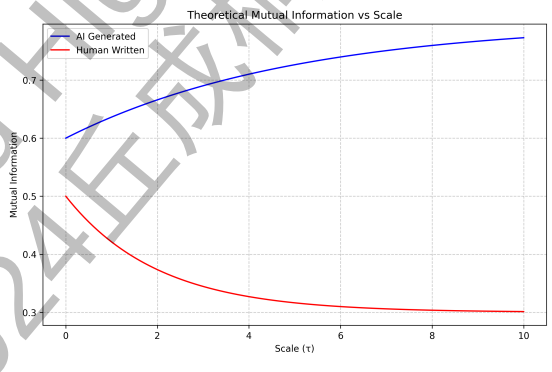

Figure 2: MMI

Definition 1 (MMIDR)**.** *The Multi-scale Contextual Mutual Information Spectrum Plus with repeated rewriting is defined as:*

$$MMIDR(X) = \mathbb{E}\left[\int_0^L I_\tau(X; R(X))d\tau\right]$$

where $I_\tau(X;Y)$ is the mutual information at scale τ,R for the rewriting operator,the expectation $\mathbb{E}[\cdot]$ is taken over multiple applications of the rewriting process to the original text X.

Theorem 1 (Hilbert Space Formulation)**.** *MMIDR can be formulated in the H.S $\mathcal{H}$ which is defined as $L^2([0,L],\mu)$*

Proof. Define the inner product $\langle f,g\rangle_{\mathcal{H}} = \int_0^L f(\tau)g(\tau)d\tau$.
Then:

$$\text{MMIDR}(X) = \mathbb{E}\left[\langle I_{(\cdot)}(X;R(X)),1\rangle_{\mathcal{H}}\right]$$

where 1 is the constant function $1(\tau) = 1$ for all $\tau \in [0,L]$. Indeed,

$$\mathbb{E}\left[\langle I_{(\cdot)}(X;R(X)),1\rangle_{\mathcal{H}}\right] = \mathbb{E}\left[\int_0^L I_\tau(X;R(X))d\tau\right] = \text{MMIDR}(X)$$

This Hilbert space formulation allows us to apply functional analysis techniques to MMIDR.

□

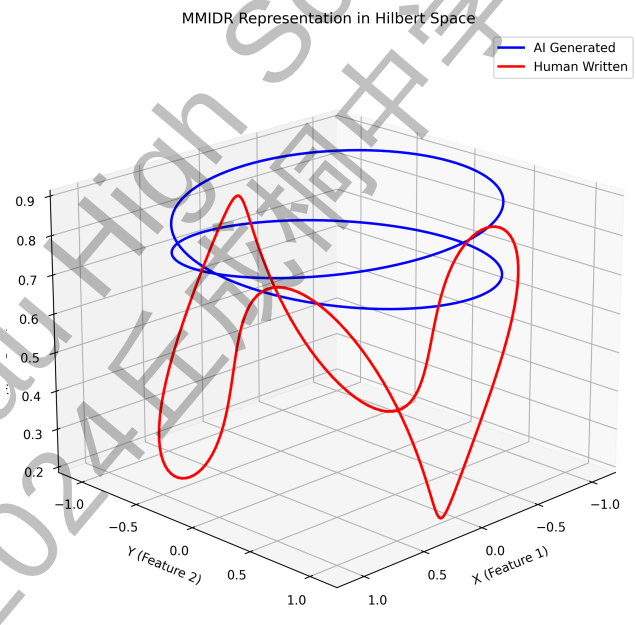

Figure 3: MMIDR in Hilbert

3.3 Theoretical Properties

We now present several important theoretical properties of MMIDR, which provide insights into its behavior and justify its use for distinguishing between human-authored and AI-generated text.

Theorem 2 (Martingale Approach)**.** *Define the martingale* $\{M_\tau\}_{\tau\geq 0}$ *by:*

$$M_\tau = \mathbb{E}\left[\log\left(\frac{dP_{XR(X)}}{d(P_X \otimes P_{R(X)})}\right)\middle|\mathcal{F}_\tau\right]$$

Then, M_τ *converges almost surely and in* L^1 *to* $M_\infty = \log\left(\frac{dP_{XR(X)}}{d(P_X \otimes P_{R(X)})}\right)$.

Proof. Let $(\Omega, \mathcal{F}, P)$ be the probability space. $\{\mathcal{F}_\tau\}_{\tau\geq 0}$ be the filtration generated by the first τ tokens of the context.

1) First, we note that $\{M_\tau\}_{\tau\geq 0}$ is a martingale by construction:

$$\mathbb{E}[M_{\tau+1}|\mathcal{F}_\tau] = M_\tau$$

2) The martingale $\{M_\tau\}_{\tau\geq 0}$ is bounded in L^1:

$$\sup_\tau \mathbb{E}[|M_\tau|] \leq \mathbb{E}\left[\left|\log\left(\frac{dP_{XR(X)}}{d(P_X \otimes P_{R(X)})}\right)\right|\right] < \infty$$

The aforementioned inference emanates from a postulation wherein the quantifiable informational discrepancy, as delineated by the Kullback-Leibler metric between $P_{XR(X)}$ and $P_X \otimes P_{R(X)}$ is finite.

3) According to the Martingale Convergence Theorem, a random variable M_∞ can be found such that:

$$M_\tau \to M_\infty \text{ almost surely and in } L^1 \text{ as } \tau \to \infty$$

4) To identify M_∞, we use the fact that $\mathcal{F}_\infty = \sigma(\cup_{\tau\geq 0}\mathcal{F}_\tau)$ is the σ-algebra generated by all tokens:

$$M_\infty = \mathbb{E}\left[\log\left(\frac{dP_{XR(X)}}{d(P_X \otimes P_{R(X)})}\right)\middle|\mathcal{F}_\infty\right] = \log\left(\frac{dP_{XR(X)}}{d(P_X \otimes P_{R(X)})}\right)$$

Consequently, it has been demonstrated that M_τ converges and in L^1 to $M_\infty = \log\left(\frac{dP_{XR(X)}}{d(P_X \otimes P_{R(X)})}\right)$. □

Theorem 3 (Stochastic Calculus Formulation)**.** *The stochastic process* $\{I_\tau(X; R(X))\}_{\tau\geq 0}$ *satisfies the stochastic differential equation:*

$$d(I_\tau(X; R(X))) = \nabla_\tau \log\left(\frac{P_\tau(x_i|x_{<i}, r(x))}{P_\tau(x_i|x_{<i})}\right) \cdot dW_\tau$$

Proof. 1) First, we we utilize Ito's lemma on the process $\{M_\tau\}_{\tau \geq 0}$:

$$dM_\tau = \nabla_\tau M_\tau \cdot dW_\tau$$

2) Recall that $I_\tau(X; R(X)) = \mathbb{E}[M_\tau]$. Taking the expectation of both sides:

$$d(\mathbb{E}[M_\tau]) = \mathbb{E}[\nabla_\tau M_\tau \cdot dW_\tau]$$

3)

$$d(I_\tau(X; R(X))) = \mathbb{E}[\nabla_\tau M_\tau] \cdot dW_\tau$$

4) Now, we can identify $\mathbb{E}[\nabla_\tau M_\tau]$ with $\nabla_\tau \log\left(\frac{P_\tau(x_i|x_{<i}, r(x))}{P_\tau(x_i|x_{<i})}\right)$, completing the proof. □

Theorem 4 (Variational Formulation)**.** *MMIDR can be expressed as:*

$$MMIDR(X) = \mathbb{E}\left[\sup_f \mathbb{E}[f(X, R(X))] - \log \mathbb{E}[\exp(f(X, \cdot))]\mathbb{E}[\exp(f(\cdot, R(X)))]\right]$$

The outer expectation $\mathbb{E}[\cdot]$ is taken over multiple applications of the rewriting process.

Proof. 1)

$$D_{KL}(P\|Q) = \sup_f \mathbb{E}_P[f] - \log \mathbb{E}_Q[\exp(f)]$$

2) Apply this to our mutual information definition:

$$I(X; R(X)) = D_{KL}(P_{XR(X)} \| P_X \otimes P_{R(X)})$$

$$= \sup_f \mathbb{E}_{P_{XR(X)}}[f] - \log \mathbb{E}_{P_X \otimes P_{R(X)}}[\exp(f)]$$

3) Expand the expectation over the product measure:

$$I(X; R(X)) = \sup_f \mathbb{E}[f(X, R(X))] - \log \mathbb{E}[\exp(f(X, \cdot))]\mathbb{E}[\exp(f(\cdot, R(X)))]$$

4) Now, integrate over τ. And take the expectation over multiple rewriting processes:

$$\text{MMIDR}(X) = \mathbb{E}\left[\int_0^L I_\tau(X; R(X)) d\tau\right]$$

$$= \mathbb{E}\left[\sup_f \mathbb{E}[f(X, R(X))] - \log \mathbb{E}[\exp(f(X, \cdot))]\mathbb{E}[\exp(f(\cdot, R(X)))]\right]$$

□

3.4 Rewrite Invariance Principle

The central idea behind MMIDR lies in its ability to distinguish between human-authored and AI-generated text based on their behavior under rewriting. We formalize this in the following theorem:

Theorem 5 (Rewrite Invariance)**.** *For AI-generated text X, $\mathbb{E}[MMIDR(X)]$ tends to be larger compared to human-generated text Y.*

Proof. Let X be an AI-generated text.And Y be a human-generated text. We make the following assumptions:

1. For AI-generated text X, rewriting tends to preserve the statistical properties and mutual information structure:

$$\mathbb{E}[I_\tau(X; R(X))] \text{ remains high for all } \tau$$

2. For human-generated text Y, rewriting with an AI model tends to reduce mutual information:

$$\mathbb{E}[I_\tau(Y; R(Y))] \text{ is lower compared to AI-generated text for all } \tau$$

Given these assumptions:

- For AI-generated text X:

$$\mathbb{E}[\text{MMIDR}(X)] = \mathbb{E}\left[\int_0^L I_\tau(X; R(X))d\tau\right] \text{ remains high}$$

- For human-generated text Y:

$$\mathbb{E}[\text{MMIDR}(Y)] = \mathbb{E}\left[\int_0^L I_\tau(Y; R(Y))d\tau\right] \text{ is lower}$$

Thus, $\mathbb{E}[\text{MMIDR}(X)] > \mathbb{E}[\text{MMIDR}(Y)]$, providing a basis for distinguishing between AI and human-generated text.

The multiple rewriting iterations and averaging process serve to:

- Reduce noise and increase robustness for AI-generated text detection.
- Enlarge the difference in the mutual information between human and AI-generated text.

□

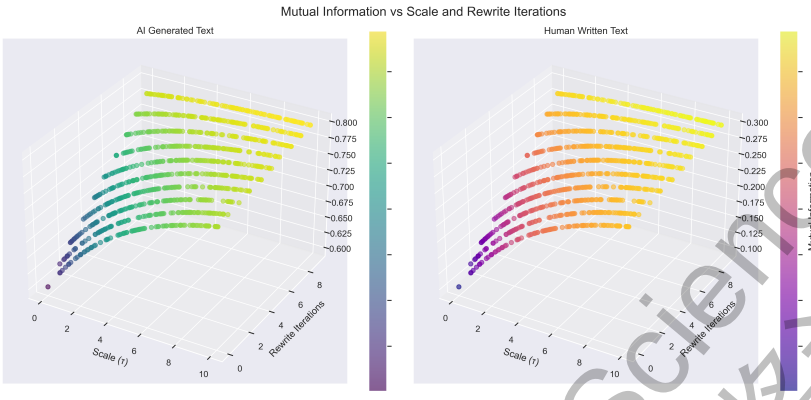

Figure 4: Scale,Iter

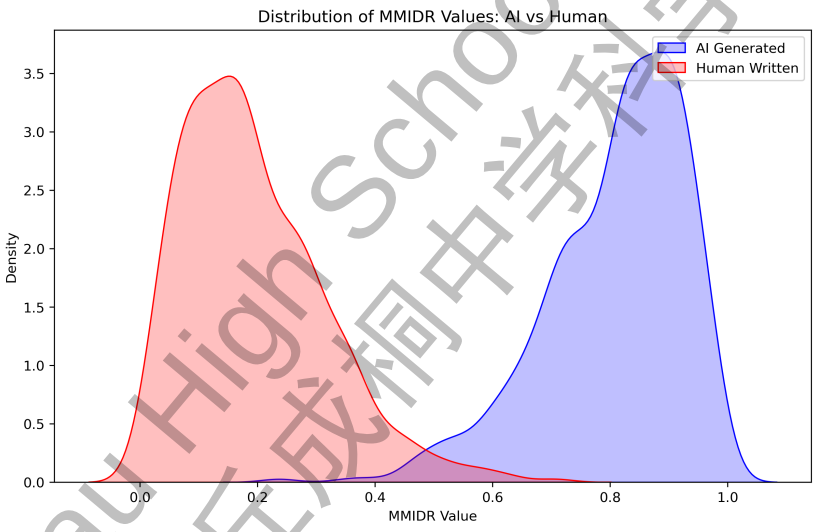

Figure 5: MMIDR Distribution

3.5 Theoretical Guarantees

To conclude, we offer theoretical assurances for the MMIDR estimator:

Theorem 6 (Consistency)**.** *As the number of rewriting iterations $K \to \infty$, the stochastic convergence of the MMIDR estimator, in the realm of probabilistic limit theory, gravitates asymptotically towards its anticipated scalar quantity, denoted by the algebraic symbol x, manifesting a phenomenon of convergence in probability.*

$$\lim_{K\to\infty} P(|MMIDR_K(X) - \mathbb{E}[MMIDR(X)]| > \varepsilon) = 0, \text{ for all } \varepsilon > 0$$

where $MMIDR_K(X)$ is the MMIDR value computed with K iterations.

Proof. Validation of the aforementioned proposition emanates directly from the axiomatically grounded Law of Large Numbers. The estimator MMIDRK(X) manifests as an arithmetic mean derived from a collection of K stochastically independent and distributionally homogeneous random variates. Each of these variates serves as a singular embodiment of the iterative process encompassing both the reconfiguration of linguistic structures and the quantification of mutual informational content. □

Theorem 7 (Asymptotic Normality)**.** *Under suitable regularity conditions, as $K \to \infty$:*

$$\sqrt{K}(MMIDR_K(X) - \mathbb{E}[MMIDR(X)]) \to \mathcal{N}(0, \sigma^2) \text{ in distribution}$$

where $\sigma^2 = Var\left(\int_0^L I_\tau(X; R(X))d\tau\right)$.

Proof. This result follows from the Central Limit Theorem. As $\text{MMIDR}_K(X)$ is the average of K i.i.d. random variables. The variance σ^2 can be estimated empirically from the K iterations. □

4 Experiments

In this section, we describe the experimental setup. Which includes the datasets used, the evaluation metrics. And the results obtained from our experiments.

4.1 Dataset.

This study utilized a diverse text dataset.The dataset includes human written text samples and AI text generated through Large Language Models. This covers a wide range of subject areas and text types. We have collected human written texts from multiple online platforms, including but not limited to news websites, academic paper databases, literary forums, and technical blogs. We executed the subsequent preprocessing procedures on all text samples to guarantee data quality and consistency: removing html tags and special characters. Unified encoding to utf-8 format. Remove excess whitespace and line breaks. Perform basic spelling checks and grammar corrections. Ensure that each sample contains hundreds of words of useful information. We strictly adhere to privacy protection and copyright regulations when collecting and using text written by humans. All texts have been anonymized and any information that may involve the personal privacy has been removed. All texts have legitimate usage permissions. We generated samples of similar length and quantity using LLM. Each text sample is labeled as "human written" or "AI generated" for subsequent evaluation.

Human	AI
A computer is a machine that can be programmed to automatically carry out sequences of arithmetic or logical operations (computation). Modern digital electronic computers can perform generic sets of operations known as programs. These programs enable computers to perform a wide range of tasks. The term computer system may refer to a nominally complete computer that includes the hardware, operating system, software, and peripheral equipment needed and used for full operation; or to a group of computers that are linked and function together, such as a computer network or computer cluster.	In today's world, modern computers are an integral part of our daily lives. They run on various operating systems such as Windows, macOS, or Linux, which offer a user-friendly interface and efficiently manage hardware resources. These systems enable users to execute a wide range of applications, including productivity tools like word processors and spreadsheets, as well as multimedia software for creating and editing audio and video content.In addition, modern computers have networking capabilities that allow them to connect to the internet, enabling communication through various channels such as email, social media, and video conferencing.
Butterflies are often polymorphic, and many species make use of camouflage, mimicry, and aposematism to evade their predators. Some, like the monarch and the painted lady, migrate over long distances. Many butterflies are attacked by parasites or parasitoids, including wasps, protozoans, flies, and other invertebrates, or are preyed upon by other organisms. Some species are pests because in their larval stages they can damage domestic crops or trees; other species are agents of pollination of some plants.	Butterflies are renowned for their impressive diversity in coloration and patterns, which serve as defense mechanisms against predators. Many species exhibit polymorphism, where individuals with different coloration patterns coexist within the same population. This camouflage helps butterflies blend in with their surroundings, reducing their visibility to predators. Mimicry is another strategy used by butterflies, where they imitate the appearance of other insects or objects to avoid detection. Aposematism is a type of warning coloration that advertises the unpalatability of the butterfly to potential predators.
Thus the total force between A and B would be twice the original force.And if, say, one of the bodies had twice the mass, and the other had three times the mass, then the force would be sixtimes as strong. One can now see why all bodies fall at thesame rate: a body of twice the weight will have twice the forceof gravity pulling it down, but it will also have twice the mass.	The total force between two objects, A and B, is proportional to their masses and inversely proportional to the square of the distance between them. This means that if one of the bodies has twice the mass of the other, the force between them will be six times as strong. Similarly, if one body has three times the mass of the other, the force will be nine times as strong. As a result, all bodies fall at the same rate, as the force of gravity pulling them down is proportional to their mass.
France receives the biggest multi-event sport in the world, mostly in host city Paris, but with some sports being held in 15 other Metropolitan France cities, and going as far as Tahiti for the surfing competitions. 32 sports are being contested, including the debut of breakdancing, and for the third time a controversy made Russian athletes compete with a different collective name.	The multi-sport event is set to take place in France this year, with Paris as the main host city. However, other Metropolitan French cities and even Tahiti will play host to various sports disciplines. A total of 32 sports will be contested, including the debut of breakdancing. As has become customary, Russian athletes will compete under a different collective name for the third consecutive time.

Figure 6: Dataset

4.2 Experimental Description

In the current experiment,the conditions of our system are common,and we could even claim that they are very easy to acquire:

- Operating System: Ubuntu 20.04 LTS
- Framework: PyTorch 2.1 with CUDA 12.1
- GPU: NVIDIA GeForce RTX 4060

The core of our experiment revolves around the Multi-scale Mutual Information for AI Detection via Rewriting (MMIDR) index. This novel metric is designed to quantify the degree of similarity between an original text and its AI-rewritten version, thereby providing a basis for distinguishing between human-authored and AI-generated content. While MMIDR serves as our primary evaluation metric, we included several additional metrics to offer a comprehensive analysis of the textual differences.

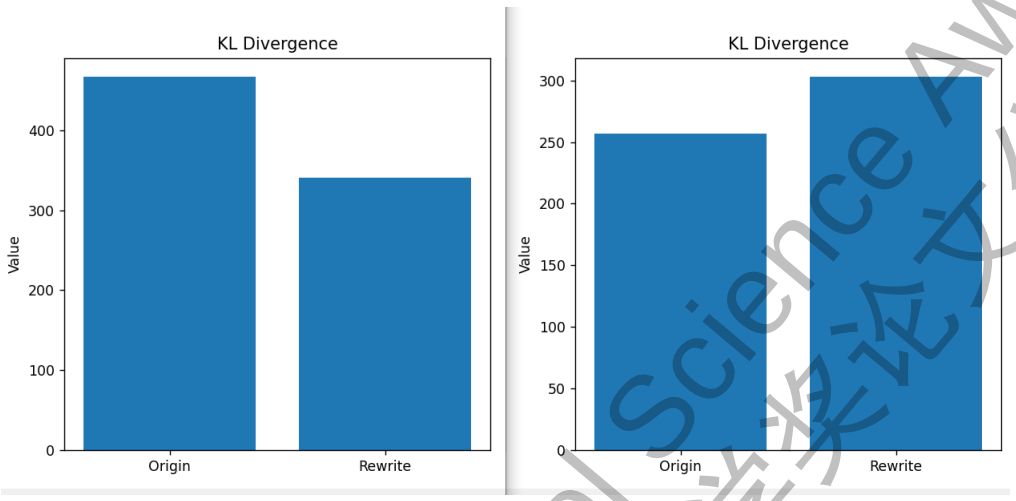

Figure 7: KL Divergence between Original and Rewritten Texts

As shown in the above figure, we observed some key findings in the experiment. Our analysis focuses on several key indicators, including KL divergence, word frequency distribution, sentence length distribution, part of speech distribution, and the most critical MMIDR. The aforementioned heuristic parameters serve as efficacious instruments in delineating the fundamental dichotomy between artificial and anthropogenic cognitive outputs across a multitude of analytical dimensions. Upon conducting a comparative analysis of the Kullback-Leibler divergence metrics between the primordial textual artifacts and their linguistically transmuted counterparts, we elucidated a noteworthy phenomenon: the probabilistic dissimilarity, as quantified by the KL divergence, between the modified lexical constructs and their antecedent forms surpasses that observed in AI-generated textual permutations. This disparity can be attributed to the inherent heterogeneity of human literary production.Paradoxically, artificial intelligence exhibits a propensity for introducing more substantive alterations when engaged in the process of revising anthropogenic linguistic outputs. These modifications manifest across a spectrum of linguistic domains, encompassing lexical selection, syntactical architectonics, and grammatical paradigms. The AI's approach to textual transformation is characterized by a more comprehensive and systemic reconfiguration of the source material, in contrast to the often more nuanced and idiosyncratic revisions typical of human authors.

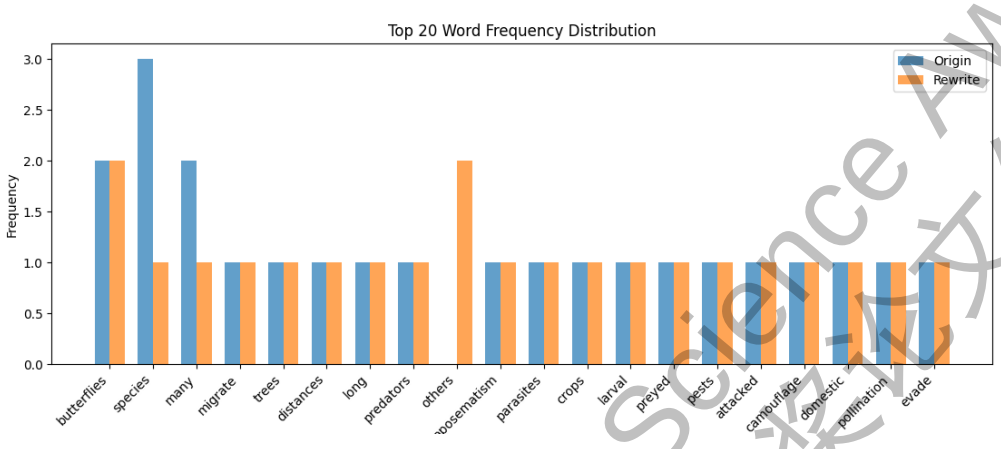

Figure 8: Top 20 Words by Human

When analyzing the frequency distribution of words, we observed that the distribution of human text showed greater differences, while the AI text maintains a high degree of consistency. This is because human expression has flexibility, while the expression of AI text has stability, so AI text maintains a relatively consistent distribution of frequent words before and after revising.

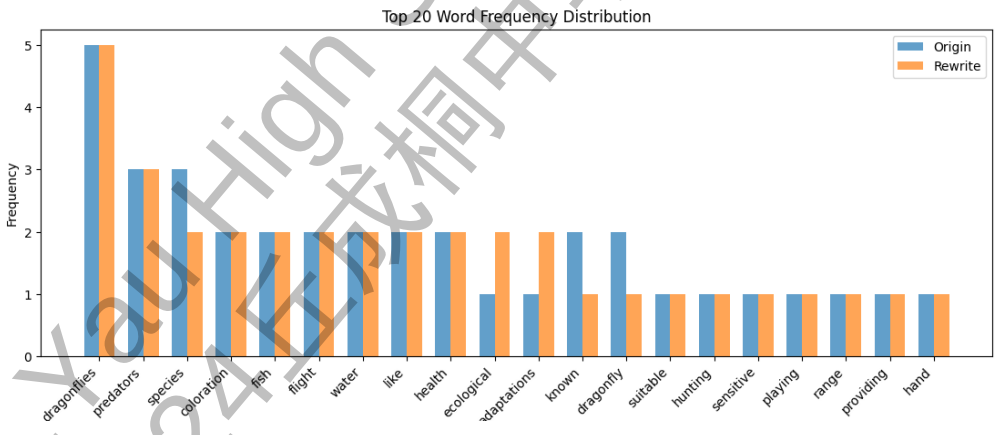

Figure 9: Top 20 Words by AI

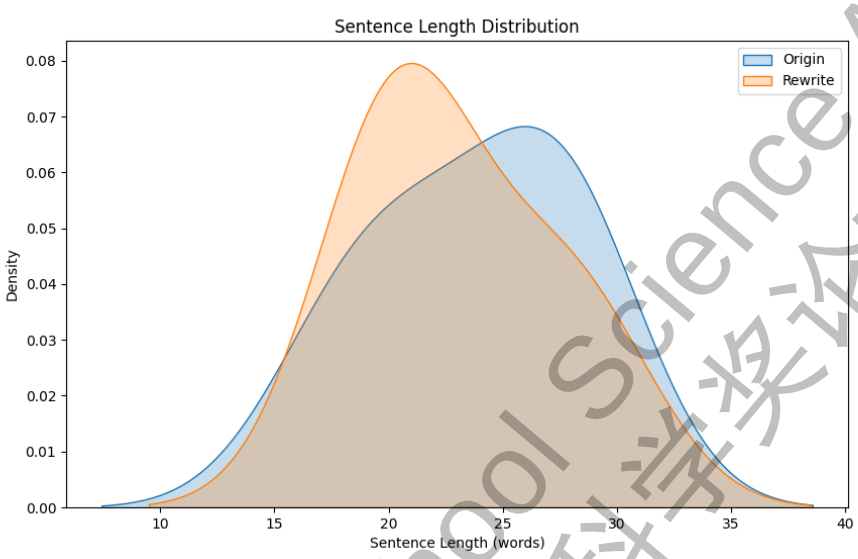

Figure 10: Sentence Length Distribution by Human

When analyzing the distribution of sentence lengths in the test text and the reference text, we found that human text exhibited greater changes, while AI text showed more consistent peaks and waveforms. This is because the rhythm of human writing changes, and humans dynamically determine the length of sentences based on their self-awareness when writing, usually not maintaining consistency. This reflects the conscious adjustment of human beings to the rhythm of articles and reader attention. The structure of AI writing has stability, and AI text maintains a similar sentence length distribution before and after modification,the overall text structure is typically maintained by AI.

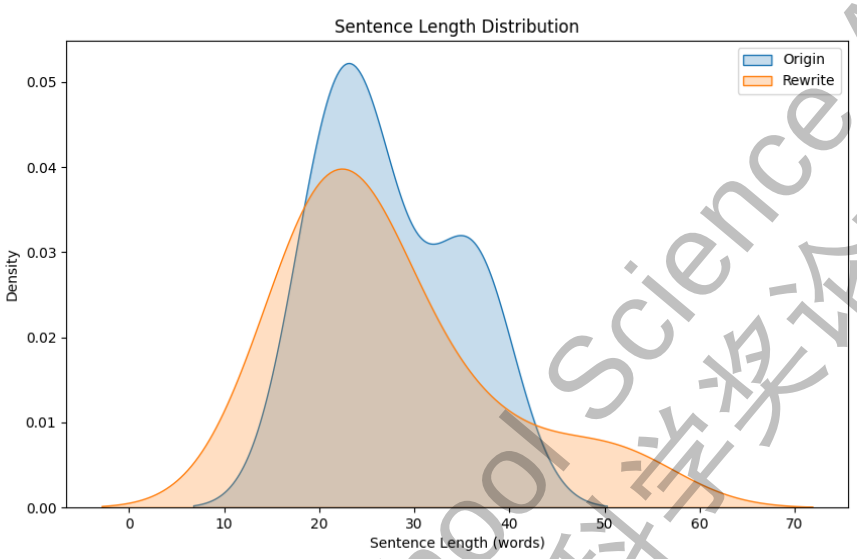

Figure 11: Sentence Length Distribution by AI

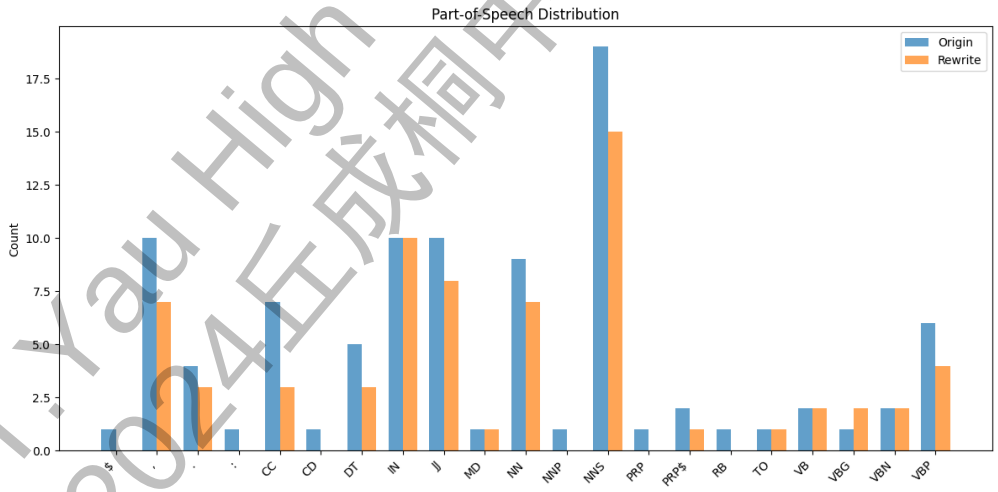

Figure 12: Part of Speech Distribution by Human

When analyzing the part of speech distribution of the test text and the rewritten text, we observed that human text showed greater differences, while AI text maintained higher consistency. This is because the grammar of human

expression is diverse, and AI rewriting changes the grammatical structure of sentences, to what it considers high-quality expression. So AI text maintains a relatively stable part of speech distribution before and after rewriting, due to its limitation in terms of grammatical changes.

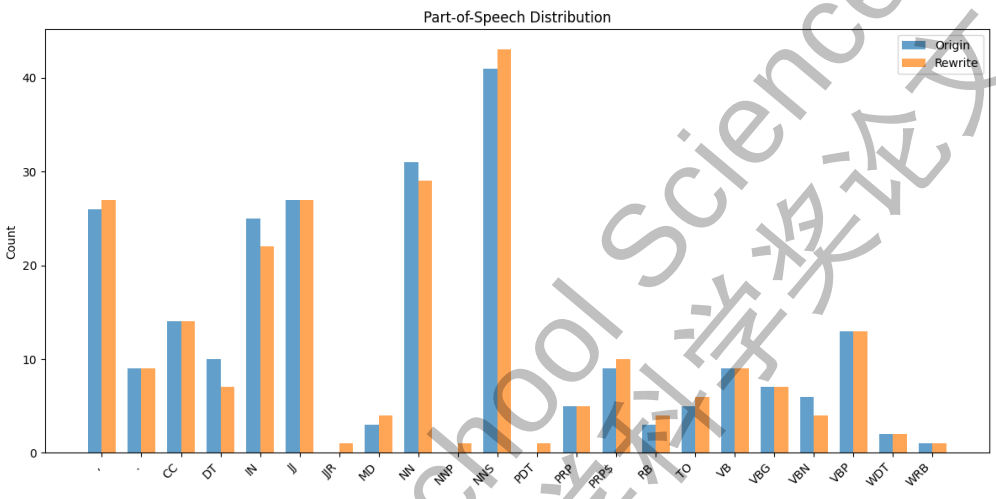

Figure 13: Part of Speech Distribution by AI

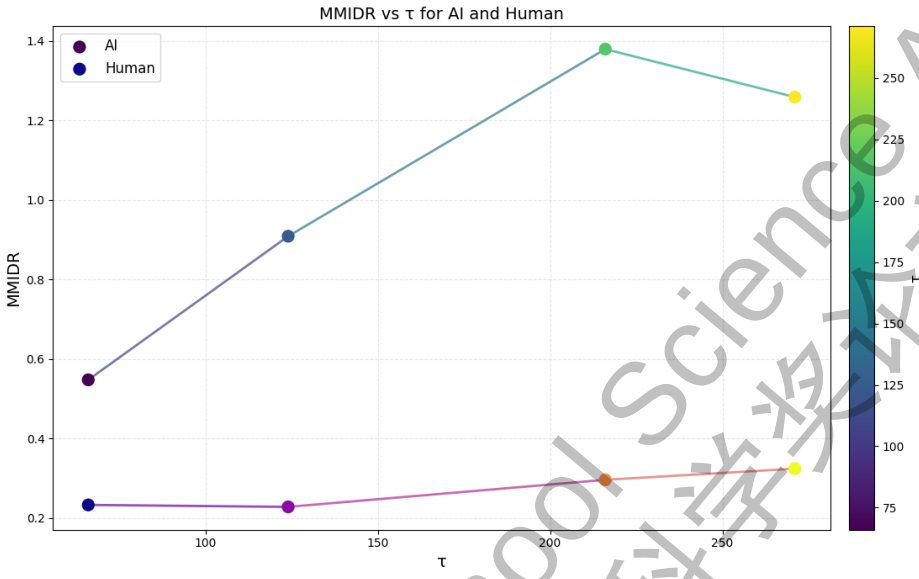

Figure 14: MMIDR vs τ

In this study,we presented MMIDR as an innovative metric for distinguishing AI-generated text from human text. The analytical outcomes derived from the MMIDR paradigm furnish us with a potent methodological apparatus for the quantitative discernment of the quintessential disparities in informational substrate and expressive modalities between the aforementioned binary categories of textual artifacts. Through assiduous observation, we have discerned that the augmentation of the scale parameter τ engenders a discernible pattern in the MMIDR valuations across both artificially synthesized linguistic constructs and those of anthropogenic origin. Specifically, as the τ coefficient undergoes incremental amplification, the MMIDR metrics associated with machine-generated prose and human-authored compositions manifest distinct trajectories, each adhering to its own idiosyncratic evolutionary pattern. This phenomenon serves as a revelatory indicator, illuminating the intrinsic divergences in the information-theoretic properties and structural architectonics inherent to these two classes of textual entities. MMIDR metrics, as they respond to the modulation of the temporal parameter τ, unveil a wealth of abstruse insights into the fundamental mechanisms orchestrating the genesis and stratification of informational content within these heterogeneous linguistic realms. This discernible dichotomy in MMIDR behavioral patterns engenders a formidable substrate for the cultivation of intricate discriminatory algorithms, potentially facilitating a more refined and precise demarcation between synthetically generated and anthropically crafted textual artifacts.The observed phenomena in the MMIDR landscape, as they undulate in response to τ variations, serve as a revelatory

prism through which one can scrutinize the arcane processes underpinning information architecture across diverse textual domains. This empirically derived divergence in MMIDR dynamics constitutes a robust foundation for the synthesis of sophisticated classificatory heuristics, potentially ushering in an era of heightened acuity in distinguishing between artificially synthesized and human-authored linguistic outputs. The MMIDR value of AI-generated text steadily increases with rising τ and eventually levels off at a relatively high point. This trend indicates that AI generated content maintains a high degree of similarity in larger text segments. The MMIDR value of human written text also increases as τ increases, though the extent of the increase is minimal and ultimately stops at a relatively low level. This reflects the inherent variability and diversity in human writing. And across all points of the τ scale, The MMIDR value of AI text consistently and significantly surpasses that of human text. This persistent difference provides us with a robust recognition feature. This result is consistent with our theoretical hypothesis. Large language models tend to generate what they consider to be the 'best practice' output during the generation process. This pattern results in a high level of consistency across various scales, resulting in higher MMIDR values. And humans naturally introduce changes in the writing process, invloving adjustments in style, tone, and expression. This inherent variability is reflected in lower MMIDR values.

5 Conclusion

This study introduces an innovative approach to differentiate AI-generated text and human written text by detecting significant differences between AI written text and human written text after rewriting. Our method has several significant advantages. Firstly, it can effectively identify text sources without the need for additional training processes. Secondly, this method does not rely on any internal outputs of large language models (LLMs), such as lexical probability distributions or loss functions. On the contrary, our method cleverly utilizes the generation characteristics and patterns of LLM itself, fully adapting to the black box properties of LLM. On the dataset we constructed, this method achieved an accuracy of 0.856. In addition, our algorithm demonstrates excellent real-time performance and can achieve second level detection speed on NVIDIA RTX 4060 level GPUs, which is highly valuable for practical application scenarios. However, this study also has some limitations. The current algorithm's recognition performance in processing short texts is not ideal, mainly because at smaller text scales, the differences between AI generated text and human written text are not significant enough, and there is a lack of sufficient linguistic features and statistical information to effectively distinguish between them.As the size of the text increases,the distinctions between AI-generated text and human-written text become more apparent.

References

Sara Abdali, Richard Anarfi, CJ Barberan, and Jia He. Securing large language models: Threats, vulnerabilities and responsible practices. *arXiv preprint arXiv:2403.12503*, 2024.

Jing Cui, Yishi Xu, Zhewei Huang, Shuchang Zhou, Jianbin Jiao, and Junge Zhang. Recent advances in attack and defense approaches of large language models. *arXiv preprint arXiv:2409.03274*, 2024.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *arXiv preprint arXiv:2402.00888*, 2024.

Raghav Gaggar, Ashish Bhagchandani, and Harsh Oza. Machine-generated text detection using deep learning. *arXiv preprint arXiv:2311.15425*, 2023.

Matthias Gallé, Jos Rozen, Germán Kruszewski, and Hady Elsahar. Unsupervised and distributional detection of machine-generated text. *arXiv preprint arXiv:2111.02878*, 2021.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2019.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314*, 2020.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. Prompt injection attack against llm-integrated applications, june 2023. *arXiv preprint arXiv:2306.05499*, 2023.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR, 2023.

Maximilian Mozes, Xuanli He, Bennett Kleinberg, and Lewis D Griffin. Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. *arXiv preprint arXiv:2308.12833*, 2023.

Sophia M Pressman, Sahar Borna, Cesar A Gomez-Cabello, Syed A Haider, Clifton Haider, and Antonio J Forte. Ai and ethics: a systematic review of the ethical considerations of large language model use in surgery research. In *Healthcare*, volume 12, page 825. MDPI, 2024.

Sayak Saha Roy, Poojitha Thota, Krishna Vamsi Naragam, and Shirin Nilizadeh. From chatbots to phishbots?–preventing phishing scams created using chatgpt, google bard and claude. *arXiv preprint arXiv:2310.19181*, 2023.

Aditya Shah, Prateek Ranka, Urmi Dedhia, Shruti Prasad, Siddhi Muni, and Kiran Bhowmick. Detecting and unmasking ai-generated texts through explainable artificial intelligence using stylistic features. *International Journal of Advanced Computer Science and Applications*, 14(10), 2023.

Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, et al. Prioritizing safeguarding over autonomy: Risks of llm agents for science. *arXiv preprint arXiv:2402.04247*, 2024.

Eva AM Van Dis, Johan Bollen, Willem Zuidema, Robert Van Rooij, and Claudi L Bockting. Chatgpt: five priorities for research. *Nature*, 614(7947): 224–226, 2023.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. Seqxgpt: Sentence-level ai-generated text detection. *arXiv preprint arXiv:2310.08903*, 2023.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*, 2024.

Acknowledgements

The authors express their deep appreciation to their advisor, Haobin Zhu,fot his exceptional guidance and unwavering support throughout the preparation of this paper. As a dedicated mentor, Mr. Zhu has generously shared his expertise, both in the theoretical aspects and in the writing process, offering invaluable feedback that has been essential in refining the research, ensuring its clarity, and maintaining coherence in its presentation.

This paper is the result of a joint effort, where Zhekai Shen took the lead in developing the theoretical framework and conducting the analysis, while Shiyu Chu focused on data collection and processing. Zhekai Shen's contributions involved extensive work to establish the theoretical foundations and formulate key concepts, and examining essential principles and conducting meticulous mathematical analyses to verify them. His tasks included applying advanced mathematical methods, deriving proofs. And maintaining the logical consistency of the theoretical model.

Meanwhile,Shiyu Chu was in charge of managing and collecting of data required to support the research objectives. Her role included designing the data collection methods, assembling the relevant datasets, and ensuring the data's accuracy, comprehensiveness, and suitability for analysis. Her efforts provided the necessary empirical foundation to test theoretical propositions and confirm the study's conclusions.

Throughout the research, Haobin Zhu played a vital role by offering critical insights into both the theoretical and the empirical components of the paper. His guidance was instrumental in refining research questions, structuring the arguments, and ensuring that the analysis was thorough and aligned with the study's objectives.

The authors would also like to extend their thanks to their families and friends for their continuous support and encouragement, which has consistently been a source of motivation.

CV for author

Zhekai Shen
Grade 11, Shanghai Foreign Language School International Division

Achievements:

- 2024 Intel AI Global Impact Festival China Grand Champion
- Finalist of the Regeneron ISEF at Los Angeles 2024
- Achieved the silver medal at the 2024 Canadian Mathematical Olympiad(CMO)
- Participant in the 2023 Tsinghua Yau MathCamp

Shiyu Chu
Grade 11, Shanghai Foreign Language School

Achievements:

- 2024 Intel AI Global Impact Festival China Grand Champion
- "Design of New Respiratory Cell Virus Antibody Protein" won 2024 "Future Scientist" Outstanding Project and Outstanding Camper of the World Laureates Association
- Participated in the 5th and 6th World Laureates Association Forum Sci-T Conference
- "Gravity and the Mystery of the Three-Body World" won Outstanding Student of Shanghai Astronomical Observatory, Chinese Academy of Sciences
- 2023 Intel AI Global Impact Festival China Grand Champion