Information Entropy for LLM-generated Text Detection

Xiaoquan Yi[1], **Haozhao Wang**[1*], **Jingcai Guo**[2*]

[1]School of Computer Science and Technology, Huazhong University of Science and Technology

[2]Department of Computin/LSGI, The Hong Kong Polytechnic University, Hong Hong SAR.

{yixiaoquan, haozhaowang}@hust.edu.cn, jingcai.guo@gmail.com

Abstract

With the explosive growth of text generated by large language models (LLMs) on the internet, technologies for detecting LLM-generated text are increasingly important. However, existing technologies often require frequent access to LLMs during detection, significantly restricting their application in scenarios where only the text is accessible. To tackle this challenge, we identify that human-written texts exhibit higher entropy than LLM-generated texts. Motivated by this, we propose a novel method named IED, which leverages the **I**nformation **E**ntropy for generative text **D**etection. Specifically, IED first transforms the given text into a vector of which each dimension represents the *information entropy* of each word, and then adopts a classifier to conduct the detection. Further, we propose IEGD which adopts the *information gain* to construct the vector by enhancing the differentiation between human-written and LLM-generated texts. To evaluate the effective of the proposed method, we construct a new large-scale dataset comprised of generated texts using various LLMs (LLaMA-7B, GPT-NeoX, etc.). Extensive experiments show that the proposed methods achieve improvement over state-of-the-art methods by up to 10.81%.

1 Introduction

Large language models (LLMs) recently exhibited the powerful ability to generate textual data [Liu *et al.*, 2023b; Li *et al.*, 2024a; Li *et al.*, 2025b], achieving great attention. However, LLMs may also be abused by malicious actors to generate large volumes of spam text, e.g., generating misleading public opinion content and fraudulent academic material [Wang *et al.*, 2025b; Li *et al.*, 2024b]. Considering this, there is an imperative for the development of efficient detection technologies capable of accurately discerning whether the text is generated by humans or LLMs.

Many approaches [Quidwai *et al.*, 2023; Bhattacharjee *et al.*, 2024; Mitchell *et al.*, 2023; Bao *et al.*, 2023] have been proposed to detect generated text, which can be generally divided into two classes depending on the type of internal information utilized by LLMs, i.e., weights and intermediate layer output. The methods utilizing weight information explore the magnitude of weight fluctuations to discern if the text has been generated by LLMs, i.e., substantial weight alterations pointing towards LLM-generated text, while slight modifications in weights are human-written [Bakhtin *et al.*, 2019]. Other methods use intermediate layer output, such as log probability for each word, to perturb the text and observe the changes in log probability [Taguchi *et al.*, 2024; Wang *et al.*, 2023b; Mitchell *et al.*, 2023; Wang *et al.*, 2023a]. A decrease in log probability indicates text generated by LLMs, while an increase or a small change in log probability suggests text written by humans. While these techniques are highly effective for detecting text when internal model information (e.g., gradients, hidden states) from LLMs is accessible, their utility is limited in black-box settings where only the generated text is available.

In this work, we seek to develop a technique that relies not on internal mechanisms of LLMs but entirely on the informational content of the text itself to determine whether text is authored by humans or generated by LLMs. Particularly, we identify that *human-written text exhibits higher information entropy (IE) than LLM-generated text*. Inspired by this observation, we introduce an innovative methodology named IED, which leverages the information entropy for generative text detection. IED first calculates the information entropy value of each word and then transforms each given input text into an IE vector by using the calculated IE. Finally, IED inputs the transformed IE vector into a binary classifier. Furthermore, we propose a novel method named IEGD, which leverages the *information gain* to construct a text vector, thereby enhancing the differentiation between human-written and LLM-generated texts. Experimental results on various datasets and models demonstrate that our method can significantly enhance detection accuracy. Our contributions are:

- We identify that human-written texts usually exhibit higher information entropy than LLM-generated text. As we know, *this research is the first to explore the detection of LLM-generated text using information entropy gain from an information-theoretic standpoint.*
- We propose the novel information entropy-based meth-

[*]Haozhao Wang and Jingcai Guo are corresponding authors.

ods for distinguishing between human and LLM-generated texts. *Our information entropy detection method is a technique that functions independently of the internal specifics (e.g., architecture, weights) of LLMs*, making it broadly applicable.

- We construct a new large-scale dataset comprised of generated texts using various LLMs and prompts from real datasets. *Extensive experiments on 9 advanced and commercial LLMs (GPT-4, Qwen-2, GPT-3.5-turbo, LLAMA-13B, etc.) show that our method outperforms state-of-the-art detection methods by up to* 10.81%.

2 Related work

In this section, we discuss representative approaches to determine whether humans or LLMs produce the text. These approaches mostly consist of leveraging internal mechanisms of LLMs, statistical analyses of feature information, and watermarking techniques.

Leveraging internal mechanisms of LLMs. Many approaches capitalize on the intricate workings of LLMs, including intermediate layer outputs and weight parameters, to discriminate between human-written and LLM-generated texts [Taguchi *et al.*, 2024; Wang *et al.*, 2023b; Wang *et al.*, 2023a; Bakhtin *et al.*, 2019]. These methods introduce subtle text modifications to track changes in log probabilities. A conspicuous decline in log probability typically points to text generated by large language models. Conversely, an elevation in log probability or insignificant variations tends to imply human authorship [Mitchell *et al.*, 2023; Bao *et al.*, 2023]. Other methods utilize weight information to explore the magnitude of weight fluctuations [Bakhtin *et al.*, 2019; Wang *et al.*, 2026]. While these methodologies demonstrate considerable efficacy in identifying text origins when afforded access to the internal workings of LLMs, their utility is constrained in scenarios where solely the text itself is accessible. Our approach differs fundamentally from theirs in both methodology and underlying assumptions. It is independent of the complex internal mechanics of LLMs and harnesses the metric of IE gain to distinguish human-written and LLM-generated text. Also, our approach's model-agnostic property confers notable universality, seamlessly detecting across a wide array of LLMs without necessitating bespoke optimization for each model [Wang *et al.*, 2025a].

Statistical analyses of feature information. Aiming at detecting text generated by small models, multiple methods employ statistical analysis to examine various features of the text, such as word choice, sentence structure, and stylistic elements, comparing them against known human or generation patterns to detect discrepancies that might indicate non-human authorship [Wang *et al.*, 2023b; Shi *et al.*, 2024; Gao *et al.*, 2018; Li *et al.*, 2025a]. Nevertheless, statistical approaches are inherently limited in detecting the text generated by LLMs. Particularly, they rely solely on patterns within vocabulary and structure, which can lead to overfitting in high-dimensional [Wu *et al.*, 2023]. As LLMs continue to advance, their capabilities have reached near parity with human performance, thereby progressively undermining the effectiveness of conventional statistical methods.

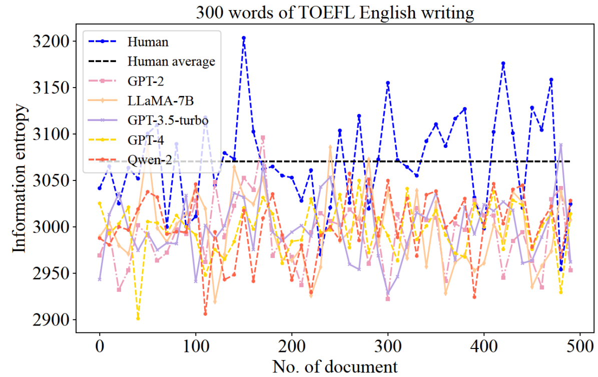

Figure 1: The chart displays the information entropy distribution of samples from the TOEFL dataset. The human-written text exhibits significantly higher information entropy compared to the text generated by LLMs.

Watermarking techniques. Watermarking constitutes an intricate, embedded detection technique that entails the sophisticated embedding of cryptic identification information within the very substrate of the generated text [Gehrmann *et al.*, 2019; Fröhling and Zubiaga, 2021; Antoun *et al.*, 2023]. It allows for identifying the source and validity of the content during subsequent detection [Lancaster, 2023; Sadasivan *et al.*, 2023]. Although the watermarking methodologies assure robustness [Kirchenbauer *et al.*, 2023], they are employed at the creation stage as opposed to being utilized for subsequent detection efforts.

3 Methodology

In this section, we introduce a framework named IED, which leverages Information Entropy to discern whether the text is human-written or LLM-generated. Specifically, IED first calculates the entropy value of each word and transforms the input text into an explicit information entropy vector. Then, it adopts a classifier to conduct the detection based on this vector. As shown in Figure 5, this architecture guides the model's semantic capacity with a discriminative, theory-driven signal, enhancing the differentiation between human-written and LLM-generated texts.

3.1 Motivation

By comparing the texts written by humans and generated by LLMs, we find that *the information entropy of human-written text is higher than LLM-generated text.* Specifically, we compiled a corpus of broader sample essays written for the TOEFL exam from online sources. We employed three advanced language models – GPT-2 [Radford *et al.*, 2019] and LLaMA-7B [Touvron *et al.*, 2023] for the essay generation task, intending to conduct a thorough comparison against essays authored by humans. Let $q(x_i^j)$ be the frequency of word x_i^j occurring in the document x_i, and W be the number of different words. We then can compute the information entropy $H(x_i)$ of each input document x_i as:

$$H(x_i) = -\sum_{j}^{W} q(x_i^j) \log_2 q(x_i^j) \tag{1}$$

As shown in Figure 1, the IE of human-written texts is higher than LLM-generated texts overall. For instance, the IE of most human-written documents is higher than 3050 (indicated by the horizontal dashed line), while that of most LLM-generated documents is smaller than 3050. The main reason is higher IE generally indicates more variability, unpredictability, and randomness in the text, which corresponds to the characteristics of human writings, which are replete with a broad spectrum of lexical choices and syntactical configurations. Conversely, LLMs tend to generate texts that often conform to structured patterns, which are inherently limited by the scope of the training corpus and algorithmic logic, lacking intrinsic human creativity.

3.2 IED: Information Entropy based Detection

In this section, we introduce IED to discern whether text x originates from LLMs or humans based on the IE. IED mainly consists of two steps, i.e., calculating the information entropy to construct an entropy-word mapping table and then training the text classifier based on this table.

LLM-generated Text Dataset Construction. Due to the lack of readily available LLM-generated text databases, we collect sufficient and representative samples of generated text for each model to be evaluated. To achieve this, we first construct a diversified prompt set $C = \{c_1, c_2, \cdots, c_n\}$ from datasets such as X-sum [Narayan *et al.*, 2018] and RoFT [Dugan *et al.*, 2023] by extracting the questions existed in their samples. Then, we randomly sample a prompt c_i from the prompt set C and input it to each LLM $j \in [S]$ to generate the text $x_{(i-1)*j+i} = \text{LLM}_j(c_i)$, where S denotes the number of LLMs. Next, the generated text $x_{(i-1)*j+i}$ is added to the generation dataset T by $T = T \cup \{x_{(i-1)*j+i}\}$. This process performs M_T iterations, i.e., constructing M_T samples for the dataset T. This approach guarantees uniformity and equity in the sampling process, cementing a robust groundwork for LLMs' subsequent evaluation and analysis.

Calculating Information Entropy. The dataset D is composed of LLM-generated dataset T and human-written dataset U. we decompose all the texts in the dataset D into a 1-gram set $G = \{g_1, \ldots, g_W\}$. Then, by counting the number $r(g_i, T)$ of occurrences of the word g_i in the entire dataset T, we can calculate the frequency $q(g_i, D)$ of each word g_i in dataset D:

$$q(g_i, D) = \frac{r(g_i, D)}{\sum_{g_j \in G} r(g_j, D)} \tag{2}$$

we calculate the $H(g_i, D)$ with equation (3).

$$H(g_i, D) = -q(g_i, D) \log_2 q(g_i, D) \tag{3}$$

Finally, the IE of the word g_i is added to the dictionary $\mathcal{D}$ by $\mathcal{D} = \mathcal{D} \cup \{(g_i, H(g_i, D)\}$.

Text Classification. For ease of classification, we transform the input text into a vector composed of information entropy gain values by retrieving the mapping table. Then, this vector is adopted as the input for the classification network. This approach enhances the model's ability to discern subtle patterns and anomalies within the text, potentially improving the accuracy and reliability of the detection process.

Specifically, we refer to the mapping table $\mathcal{M}$ to get the corresponding information entropy vector e_i for the input text x_i remaining to be detected, i.e., $e_i = [\mathcal{M}(x_i^1), \mathcal{M}(x_i^2), \ldots, \mathcal{M}(x_i^W)]$. Then, we transform the information entropy vector e_i to the backbone θ_b of the neural network, obtaining the hidden feature vector $z_i = f_b(e_i, \theta_b)$. Let θ_h be the classifier parameter. The detection label is obtained by $\hat{y_i} = f_h(z_i, \theta_h)$. Let $l(\cdot)$ be the loss function. Then, our objective is to train the classifier θ_h by minimizing the average of all sample loss $L(\hat{y_i}, y_i; \theta_h)$:

$$\min_{\theta_h} L(\theta_h) =: \frac{1}{M} \sum_{i=1}^{M} l(y_i, \hat{y_i}; \theta_h) = \frac{1}{M} \sum_{i=1}^{M} l(y_i, f_h(z_i, \theta_h)) \tag{4}$$

Note that we keep the parameter θ_b unchanged during training and only train the classification layer with $\theta_h = \theta_h - \eta \nabla L(\hat{y_i}, y_i, \theta_h)$ to improve the efficiency of fine-tuning the model parameter, where η is the learning rate.

3.3 IEGD: Information Entropy Gain based Detection

Drawing upon our insightful observations, we endeavored to leverage information entropy gain as a pivotal element in our experimental design. Particularly, built upon the IE, information entropy gain quantifies the decrease in uncertainty pertinent to data categorization, further enriching the analytical depth of IE. While IED has demonstrated robust efficacy in characterizing textual patterns, the concept of IEG extends this theoretical foundation by providing a quantitative measure of distributional differences between human and machine-generated text features. Specifically, information entropy gain enhances discrimination capability when employed with information entropy metrics, offering a systematic approach to identifying distinctive linguistic patterns between the two text categories. Therefore, we further extends the IED to IEAD that uses the information entropy gain to vectorize the input texts.

Calculating Information Entropy Gain. Let D be the same dataset composed of LLM-generated dataset T and human-written dataset U. We first calculate the information entropy $H(D)$ of the whole dataset D:

$$H(D) = \sum_{g_i \in G} H(g_i, D) \tag{5}$$

We divide the dataset D into two subsets: D_{g_i} contains the text with g_i, and $D_{\neg g_i}$ doesn't contain g. Then, we calculate the conditional entropy $H(D|g_i)$:

$$H(D|g_i) = \frac{|D_{g_i}|}{|D|} H(D_{g_i}) + \frac{|D_{\neg g_i}|}{|D|} H(D_{\neg g_i}) \tag{6}$$

Next, we calculate the information gain $\Delta H(D|g_i)$ by

$$\Delta H(D, g_i) = H(D) - H(D|g_i) \tag{7}$$

Based on the information entropy gain of all words, we transform the texts into the vectors of information entropy gain. Finally, we train the classifier θ_h using the same equation (4). The algorithm and illustration of IEGD can be found in Appendix A.

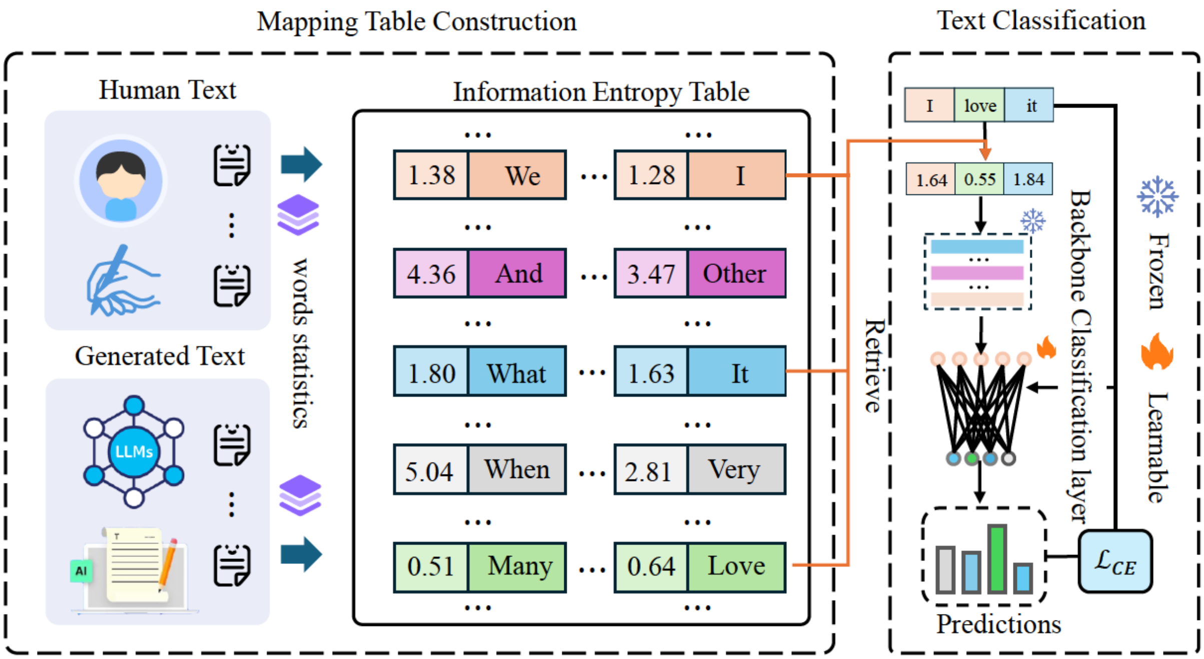

Figure 2: The detailed processes of IED. It operates through two primary stages: mapping table construction and text classification. In the first stage, we calculate per-word information entropy. In the second stage, texts are encoded into vectorial form by harnessing the entropy gain values indexed within the table, and the procedure equips the classifier with the requisite data.

4 Experiments

In this section, we first introduce the experimental setup for the datasets, implementation details of the IEGD, and the baselines. We evaluated the accuracy of our approach on four datasets and compared it with state-of-the-art algorithms. Finally, we assessed the impact of hyperparameters on our proposed method.

4.1 Experiment setting

Dataset&LLMs.Our experiment employs a meticulously curated collection of four datasets to rigorously appraise the capabilities of generated text detection across a spectrum of domains. For instance, news articles from the XSum dataset [Narayan *et al.*, 2018] are used for news detection. Wikipedia paragraphs from SQuAD contexts [Rajpurkar *et al.*, 2016] represent machine-written academic essays and prompted stories from the Reddit WritingPrompts dataset [Fan *et al.*, 2018] gauge the models' ability to generate creative writing similar to human submissions. To further test the model's robustness against shifts in data distribution. We include various topics from the RoFT dataset [Dugan *et al.*, 2023]. We use various LLMs to generate text, i.e., GPT-2 [Radford *et al.*, 2019], OPT-2.7B [Zhang *et al.*, 2022], Neo-2.7B [Gao *et al.*, 2020], LLaMA-7B [Touvron *et al.*, 2023], LLaMA-13B, GPT-NeoX [Black *et al.*, 2022], GPT-3.5-turbo, GPT-4, Qwen-2 [Wang *et al.*, 2024].

Metrics. To evaluate the detector's ability to distinguish between texts generated by LLMs or humans, we employ Accuracy (A), AUROC,F1 scores (F1) and Recall (R) as performance metrics.

Baselines. We compared our method with various approaches for detecting LLM-generated text. FastDetect-GPT [Bao *et al.*, 2023] is a method for detecting text generated by large language models (LLMs), which works by evaluating the conditional probability curvature associated with sampled alternative tokens. The **Log-Rank** is employed to assess whether text generated by a model is based on the value of the log perplexity. GLTR [Gehrmann *et al.*, 2019] integrates a suite of statistical techniques, employing visual analytics to assist users in discerning text generated by models. DetectGPT [Mitchell *et al.*, 2023] exploits the curvature characteristics of language model probability functions to discern synthesized text. LLMDet [Wu *et al.*, 2023] assesses the provenance of generated text by calculating the surrogate perplexity for each LLM. SeqXGPT [Wang *et al.*, 2023a] translates sentences into waveforms akin to those in speech processing, leveraging convolutional networks and self-attention mechanisms for sentence-level detection of generated text. GECScore [Wu *et al.*, 2024] determines whether text originates from LLMs by feeding it into a grammar error correction model, quantifying the similarity between the original and corrected texts. COCO [Liu *et al.*, 2023a] uses contrastive learning to detect text.

Implementation Details. To ensure the reproducibility and efficiency of our experiments, all computations were meticulously conducted on a high-performance computing platform running the Ubuntu operating system. The system is powered by an Intel Xeon Platinum 8176 CPU@2.10GHz, delivering robust computational capabilities for data processing. To handle the memory-intensive tasks associated with large language models, the platform is equipped with a substantial

Dataset	ACC(%)	FastDetectGPT	DetectGPT	LLMDet	SeqXGPT	GECScore	COCO	IED	IEGD
X-sum	GPT-2	72.87 ± 3.51	80.25 ± 1.58	71.07 ± 3.93	87.16 ± 1.33	82.37 ± 2.37	89.22 ± 5.11	80.81 ± 5.64	**92.60 ± 3.87**
	OPT-2.7	81.57 ± 2.37	75.47 ± 0.21	74.36 ± 4.94	84.32 ± 2.40	79.51 ± 1.03	85.25 ± 5.22	86.82 ± 2.32	**91.68 ± 2.21**
	Neo-2.7B	82.23 ± 1.67	68.80 ± 1.28	70.16 ± 2.46	79.04 ± 1.37	70.71 ± 1.57	81.86 ± 2.28	92.94 ± 3.98	**93.50 ± 3.77**
	LLaMA-7B	84.16 ± 2.61	75.30 ± 1.68	75.80 ± 1.35	85.41 ± 5.80	76.28 ± 0.58	83.93 ± 4.68	85.47 ± 1.05	**90.40 ± 0.23**
	LLAMA-13B	77.21 ± 3.62	68.80 ± 1.58	69.86 ± 3.79	74.95 ± 4.70	69.85 ± 3.12	79.28 ± 3.54	81.98 ± 1.44	**83.91 ± 3.18**
	GPT-NeoX	79.24 ± 1.07	67.71 ± 1.58	67.52 ± 4.90	76.76 ± 3.65	69.77 ± 1.82	82.38 ± 4.09	84.02 ± 5.08	**86.98 ± 1.39**
	GPT-3.5-turbo	81.58 ± 1.76	75.68 ± 2.03	65.57 ± 2.82	82.88 ± 1.40	69.99 ± 1.98	84.84 ± 8.03	87.51 ± 6.80	**92.15 ± 1.44**
	GPT-4	76.17 ± 1.81	81.89 ± 2.84	68.95 ± 4.57	84.77 ± 1.61	71.65 ± 1.34	89.62 ± 1.63	90.68 ± 2.61	**93.46 ± 1.52**
	Qwen-2	81.42 ± 2.35	70.55 ± 2.72	66.51 ± 3.73	83.87 ± 2.91	73.96 ± 1.77	87.50 ± 1.58	92.65 ± 1.68	**93.31 ± 1.24**
SQuAD	GPT-2	80.82 ± 1.95	73.78 ± 4.71	70.21 ± 1.89	89.10 ± 3.62	84.25 ± 1.68	91.47 ± 2.06	90.05 ± 3.45	**96.97 ± 1.92**
	OPT-2.7	78.58 ± 1.82	71.12 ± 1.50	72.96 ± 1.03	81.56 ± 2.37	75.67 ± 3.59	83.06 ± 2.98	86.69 ± 1.33	**87.89 ± 5.34**
	Neo-2.7B	81.69 ± 1.56	72.97 ± 2.16	78.17 ± 1.68	84.16 ± 53.5	70.45 ± 2.10	83.48 ± 1.17	76.90 ± 1.30	**86.40 ± 1.53**
	LLaMA-7B	80.99 ± 3.47	76.65 ± 0.69	78.51 ± 0.90	59.54 ± 1.91	75.06 ± 2.45	91.33 ± 4.17	89.17 ± 4.75	**95.17 ± 4.45**
	LLAMA-13B	86.91 ± 1.68	75.90 ± 4.42	73.59 ± 2.27	82.02 ± 0.19	73.73 ± 1.12	78.04 ± 0.18	82.62 ± 6.49	**87.10 ± 1.06**
	GPT-NeoX	82.11 ± 0.45	69.22 ± 2.51	67.73 ± 1.68	79.28 ± 1.84	68.87 ± 2.53	83.79 ± 1.55	80.39 ± 5.57	**85.75 ± 3.14**
	GPT-3.5-turbo	82.11 ± 0.45	78.28 ± 1.76	60.51 ± 1.96	87.08 ± 1.56	80.03 ± 2.34	87.65 ± 4.56	88.27 ± 3.17	**91.97 ± 2.93**
	GPT-4	80.17 ± 0.72	82.66 ± 1.88	60.13 ± 2.85	86.78 ± 6.87	71.21 ± 1.65	82.42 ± 3.83	88.27 ± 3.74	**98.57 ± 1.64**
	Qwen-2	81.61 ± 2.23	75.54 ± 2.39	71.46 ± 1.61	85.08 ± 1.51	65.98 ± 1.22	84.74 ± 1.49	93.11 ± 2.98	**95.32 ± 1.55**
WritingPrompts	GPT-2	78.93 ± 1.64	76.70 ± 1.93	73.62 ± 1.97	85.28 ± 4.20	83.77 ± 1.08	83.22 ± 3.08	87.27 ± 1.71	**94.32 ± 1.21**
	OPT-2.7	83.26 ± 2.97	71.33 ± 1.80	75.26 ± 1.80	78.15 ± 1.74	75.76 ± 3.90	80.81 ± 6.98	77.24 ± 3.75	**82.08 ± 7.00**
	Neo-2.7B	81.95 ± 2.00	74.07 ± 0.56	76.02 ± 3.00	80.71 ± 1.15	74.98 ± 2.22	85.10 ± 2.28	82.27 ± 1.94	**85.88 ± 0.23**
	LLaMA-7B	82.30 ± 2.55	62.97 ± 2.55	72.88 ± 3.51	73.77 ± 1.63	66.64 ± 2.65	84.48 ± 1.21	76.93 ± 1.56	**87.79 ± 1.36**
	LLAMA-13B	81.38 ± 0.26	74.88 ± 0.97	78.53 ± 3.21	86.69 ± 1.05	78.82 ± 2.20	89.33 ± 2.42	85.83 ± 1.06	**91.23 ± 5.06**
	GPT-NeoX	81.85 ± 1.45	71.04 ± 3.70	66.30 ± 2.22	80.58 ± 2.00	65.17 ± 3.48	84.19 ± 1.24	87.59 ± 4.14	**90.55 ± 1.34**
	GPT-3.5-turbo	81.85 ± 1.45	72.29 ± 3.64	67.99 ± 1.76	80.09 ± 1.52	78.92 ± 1.55	88.05 ± 3.94	89.62 ± 1.36	**89.71 ± 1.05**
	GPT-4	79.14 ± 1.24	75.26 ± 2.87	67.57 ± 1.94	86.91 ± 3.32	67.57 ± 1.94	84.52 ± 2.97	89.45 ± 1.86	**90.86 ± 1.52**
	Qwen-2	76.29 ± 3.30	65.60 ± 2.69	61.73 ± 1.73	82.33 ± 1.08	69.66 ± 1.45	85.10 ± 2.76	88.82 ± 1.26	**92.29 ± 2.13**
RoFT	GPT-2	78.87 ± 2.66	73.73 ± 2.59	68.12 ± 8.35	80.97 ± 7.99	72.92 ± 7.34	77.42 ± 5.41	80.72 ± 1.26	**82.63 ± 4.33**
	OPT-2.7	51.11 ± 5.20	76.57 ± 6.76	69.27 ± 6.53	80.58 ± 3.73	67.64 ± 2.21	69.81 ± 1.46	85.11 ± 2.60	**91.36 ± 2.70**
	Neo-2.7B	81.75 ± 2.03	70.46 ± 1.71	68.84 ± 1.78	79.42 ± 3.88	71.84 ± 5.14	66.04 ± 3.08	76.40 ± 7.76	**83.64 ± 1.21**
	LLaMA-7B	82.01 ± 1.44	57.88 ± 2.10	67.73 ± 7.47	79.81 ± 6.40	63.27 ± 1.27	72.88 ± 3.16	81.83 ± 4.98	**84.04 ± 1.60**
	LLAMA-13B	76.22 ± 1.12	78.00 ± 1.67	72.06 ± 2.04	80.97 ± 1.83	78.63 ± 2.06	83.74 ± 1.14	90.37 ± 1.93	**90.89 ± 0.88**
	GPT-NeoX	78.14 ± 1.34	72.26 ± 6.16	66.17 ± 2.22	71.46 ± 0.76	65.79 ± 0.50	70.93 ± 2.64	73.61 ± 2.91	**74.15 ± 4.70**
	GPT-3.5-turbo	81.84 ± 0.67	74.36 ± 1.81	71.48 ± 8.87	85.06 ± 1.85	66.24 ± 3.49	82.22 ± 2.49	93.16 ± 2.98	**90.19 ± 3.64**
	GPT-4	84.41 ± 1.62	75.21 ± 4.74	74.83 ± 2.68	73.51 ± 1.11	73.04 ± 1.07	82.71 ± 2.73	85.76 ± 3.05	**90.23 ± 1.21**
	Qwen-2	78.15 ± 2.46	72.77 ± 1.91	59.87 ± 3.57	88.71 ± 3.84	74.76 ± 2.24	88.54 ± 2.72	90.93 ± 4.07	**91.57 ± 1.62**

Table 1: This table compares different methods' performance on four datasets: X-sum, SQuAD, WritingPrompts, and RoFT. Our method achieves higher average accuracy across all datasets than the other methods.

503 GB of memory. To accelerate the inference and calculation of information entropy, we utilized four NVIDIA Corporation GA102GL RTX A6000 GPUs.

4.2 Performance Evaluation

Accuracy. As detailed in Table 1, our proposed method consistently demonstrates superior precision across diverse benchmarks, outperforming state-of-the-art baselines. Specifically, on the X-sum dataset for GPT-2 generated text, IEGD achieves a remarkable accuracy of 92.60%, surpassing competitive methods like DetectGPT (80.25%) and FastDetectGPT (72.87%). Beyond individual datasets, extensive experiments indicate that our method achieves an improvement over existing methods by up to 10.81%. This performance disparity validates our foundational premise: human-authored texts exhibit higher information entropy due to the complexity of human thought and rich background knowledge. In contrast, LLMs optimize for probability maximization, resulting in structured, predictable patterns with lower entropy. By quantifying this uncertainty gap, IEGD effectively transforms subtle statistical variations into discriminative signals for accurate detection.

F1-score. As illustrated in Table 2, the experimental results demonstrate the superior performance of IED and IEGD in distinguishing LLM-generated text, maintaining a robust balance between precision and recall. On the WritingPrompts dataset, IEGD achieves an F1 score of 89.97% for GPT-2 generated content, significantly outperforming baselines such as Log-Rank (47.12%) and DetectGPT (74.06%). Similarly, for OPT-2.7, IEGD attains an F1 score of 90.10%, showing a substantial lead over Log-Rank's 48.86%. This dominance extends to the RoFT dataset, where IEGD consistently surpasses competitors (e.g., 88.40% on GPT-2). By analyzing information entropy at the lexical level, our methods effectively capture the intrinsic differences in statistical distributions and underlying structures of texts. Beyond merely identifying superficial discrepancies, these entropy-based measures amplify the underlying contrast by quantitatively reflecting the varying degrees of uncertainty, randomness, and informational richness. The subtle yet crucial distinctions in textual properties between human and LLM-generated content are translated into more pronounced and empirically measurable signals, significantly enhancing detection capabilities even in

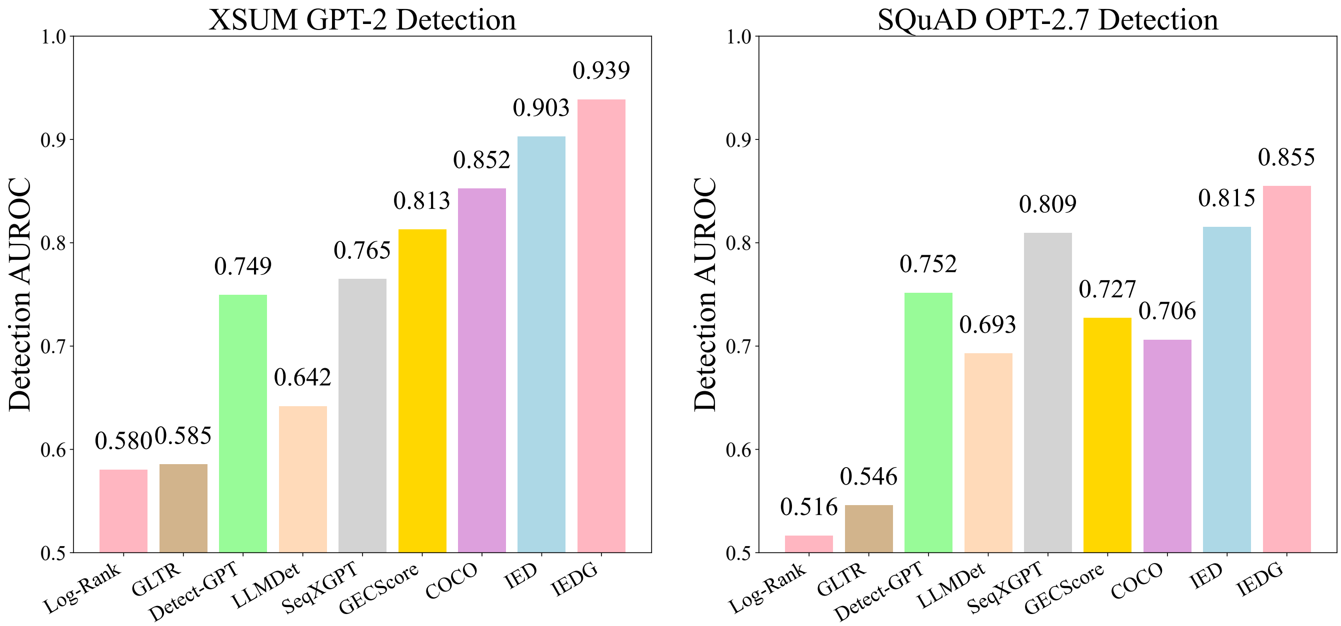

Figure 3: Performance comparison of various text generation detection methods across four detection tasks. Each bar chart displays the detection performance of different methods for a specific task, measured by AUROC. Our method performs the best across all tasks, significantly surpassing other detection methods.

Dataset	F1(%)	Log-Rank	DetectGPT	LLMDet	SeqXGPT	GECScore	COCO	IED	IEGD
WritingPrompts	GPT-2	47.12± 2.43	74.06± 1.03	66.91± 2.27	88.13± 2.53	63.35± 7.28	88.72± 4.79	85.34± 2.48	**89.97± 2.75**
	OPT-2.7	48.86± 3.31	67.43± 4.10	73.65± 3.96	85.08± 1.63	75.39± 4.68	86.30± 6.12	89.32± 2.45	**90.10± 1.95**
	Neo-2.7B	54.98± 2.04	72.46± 5.21	74.48± 1.23	79.82± 1.26	73.91± 5.55	90.79± 6.38	92.45± 1.63	**94.12± 2.11**
	LLaMA-7B	60.95± 4.18	65.89± 1.29	58.26± 3.52	76.21± 4.66	70.08± 2.43	81.93± 1.02	88.74± 3.25	**91.12± 2.22**
	LLAMA-13B	51.94± 3.48	76.35± 6.79	63.89± 3.92	76.48± 1.99	74.57± 2.45	83.07± 3.90	84.47± 3.08	**86.42± 2.93**
	GPT-NeoX	56.53± 1.10	76.62± 3.44	64.11± 10.73	80.24± 4.24	69.83± 1.48	82.16± 1.78	86.79± 2.59	**88.54± 2.22**
	GPT-3.5-turbo	50.08± 2.20	59.93± 1.71	65.39± 2.40	77.93± 2.04	63.10± 1.49	72.36± 4.79	80.54± 3.23	**89.05± 2.15**
	GPT-4	50.72± 4.66	71.92± 6.58	59.63± 1.19	85.59± 2.96	60.39± 3.15	72.91± 3.53	81.03± 2.16	**89.41± 1.22**
	Qwen-2	46.68± 3.94	67.24± 1.51	59.38± 3.65	64.87± 3.57	71.60± 1.80	76.81± 4.21	86.44± 2.04	**89.07± 2.70**
RoFT	GPT-2	53.84± 1.77	62.72± 2.39	79.66± 08	69.31± 1.38	83.42± 2.18	81.27± 3.09	87.27± 3.09	**88.40± 1.73**
	OPT-2.7	49.48± 3.04	84.28± 1.38	64.35± 1.02	79.05± 1.52	67.94± 1.35	83.41± 2.49	86.43± 1.67	**89.24± 1.80**
	Neo-2.7B	51.26± 1.88	71.96± 1.22	66.91± 1.53	81.11± 1.48	87.18± 1.80	85.16± 1.48	81.45± 1.42	**90.22± 1.25**
	LLaMA-7B	50.98± 1.36	71.41± 2.73	63.16± 2.86	80.68± 3.82	68.22± 2.27	80.41± 2.47	87.49± 1.36	**90.81± 2.49**
	LLAMA-13B	52.47± 3.48	72.47± 1.62	63.67± 1.565	78.86± 1.92	65.74± 3.15	80.79± 1.43	90.22± 1.57	**91.80± 2.91**
	GPT-NeoX	53.85± 1.58	70.83± 2.80	62.05± 1.54	77.78± 1.42	67.59± 1.93	82.72± 1.06	85.00± 1.65	**88.60± 3.36**
	GPT-3.5-turbo	50.99± 3.55	81.96± 1.02	64.71± 2.65	79.85± 1.57	68.22± 1.59	81.51± 2.10	85.16± 1.79	**89.05± 1.88**
	GPT-4	52.10± 2.37	68.07± 1.95	53.20± 1.62	78.21± 2.34	65.96± 1.35	81.62± 2.97	**87.44± 1.74**	86.85± 1.31
	Qwen-2	50.82± 2.71	71.97± 1.19	67.44± 2.11	78.98± 1.81	66.50± 1.60	80.43± 2.65	87.26± 1.54	**90.03± 2.49**

Table 2: This table compares different methods' F1-score on the WritingPrompts and RoFT dataset. Our method achieves a higher F1 score across all datasets than the other methods.

complex creative writing scenarios.

Recall. As illustrated in Table 3, the experimental results on the X-sum dataset reveal the superiority of IED and IEGD over other state-of-the-art detection methods when evaluating texts generated by various LLMs. Specifically, IEGD achieves a recall rate of 93.9% for content generated by GPT-2, significantly outperforming baselines such as Log-Rank and COCO . This trend extends to the SQuAD dataset, where both IED and IEGD consistently achieve high recall rates across a variety of models. These findings underscore the consistent advantage of IED and IEGD in identifying and distinguishing between human-authored and LLM-generated texts. The superior performance is attributed to their ability to capture both statistical and structural differences.

Ablation Study. Table 4 presents an ablation study focusing on the performance of the IED and IEGD methods when combined with different transformer-based classifier backbones. Across all datasets, the IEGD variant consistently outperforms its IED counterpart. IEGD-BERT achieves 85.99% on X-sum compared to IED-BERT's 79.56%. Similarly, IEGD-RoBERTa scores 92.41% on SQuAD, while IED-RoBERTa scores 86.91%. This suggests that using Information Entropy

Dataset	R(%)	Log-Rank	DetectGPT	LLMDet	SeqXGPT	GECScore	COCO	IED	IEGD
X-sum	GPT-2	50.58± 6.62	72.46± 1.10	69.36± 8.14	92.62± 3.03	70.42± 3.41	84.44± 4.37	90.73± 2.97	**93.17± 4.05**
	OPT-2.7	60.69± 4.97	78.54± 3.52	73.89± 1.22	81.39± 2.72	79.27± 1.96	90.33± 1.44	89.64± 2.25	**94.40± 1.44**
	Neo-2.7B	56.29± 2.48	70.05± 2.22	76.38± 1.59	85.59± 5.61	69.33± 1.14	86.84± 7.02	90.48± 3.17	**93.95± 2.88**
	LLaMA-7B	54.64± 2.25	65.64± 5.93	74.92± 1.28	81.23± 1.33	75.08± 9.48	90.00± 5.05	92.09± 2.68	**92.28± 1.23**
	LLAMA-13B	52.63± 1.07	66.12± 5.56	71.81± 3.33	75.21± 1.75	75.40± 4.52	85.96± 2.48	83.29± 1.81	**86.48± 3.24**
	GPT-NeoX	45.04± 1.73	67.48± 1.56	65.67± 2.79	75.45± 5.65	71.38± 1.19	82.05± 5.38	85.49± 1.62	**88.22± 0.44**
	GPT-3.5-turbo	47.10± 3.63	70.24± 2.25	62.54± 1.79	74.90± 2.46	65.44± 1.79	65.44± 1.71	78.70± 1.03	**82.32± 2.75**
	GPT-4	45.96± 2.37	69.15± 4.75	53.86± 1.66	75.42± 2.43	58.93± 2.08	78.59± 3.34	85.91± 1.67	**89.35± 2.42**
	Qwen-2	43.80± 2.89	66.96± 1.73	52.78± 4.42	65.48± 5.44	56.90± 2.28	70.99± 2.64	80.34± 8.59	**85.84± 2.32**
SQuAD	GPT-2	49.41± 2.05	66.48± 1.03	60.62± 1.33	74.78± 1.45	68.19± 2.27	78.94± 1.46	85.91± 2.49	**88.19± 1.89**
	OPT-2.7	56.80± 2.92	67.01± 1.02	63.62± 1.09	80.93± 2.47	63.55± 2.31	78.50± 1.95	85.40± 1.94	**90.37± 2.62**
	Neo-2.7B	48.85± 1.76	58.89± 1.93	80.40± 1.49	57.88± 1.29	78.04± 2.97	89.78± 1.14	86.01± 1.02	**90.04± 0.78**
	LLaMA-7B	48.32± 2.67	65.14± 2.16	53.60± 1.35	78.62± 1.12	64.56± 1.10	82.28± 1.33	**90.62± 1.58**	87.00± 1.49
	LLAMA-13B	51.09± 2.10	64.78± 1.09	58.70± 2.41	75.85± 1.63	66.25± 1.08	71.43± 2.75	**83.35± 1.03**	83.29± 2.56
	GPT-NeoX	46.42± 3.92	68.60± 1.88	59.51± 2.52	77.69± 1.70	65.15± 2.82	73.79± 1.23	81.87± 1.80	**84.31± 2.9**
	GPT-3.5-turbo	43.56± 1.35	73.69± 2.13	68.34± 2.22	79.03± 4.01	60.00± 2.39	75.33± 1.72	80.75± 1.93	**85.10± 1.71**
	GPT-4	50.92± 4.19	62.06± 2.45	73.29± 2.33	63.90± 2.94	77.90± 2.05	80.13± 2.93	79.13± 2.93	**86.68± 2.74**
	Qwen-2	46.23± 1.76	70.32± 3.64	61.05± 1.84	76.59± 2.63	64.41± 3.09	79.43± 2.14	81.13± 1.68	**82.07± 1.31**

Table 3: This table compares different methods' recall on the X-sum and SQuAD dataset. Our method achieves higher recall across all datasets than the other methods.

Classifier	X-sum	SQuAD	WritingPrompts	RoFT
IED-BERT	79.56 ± 3.47	79.39 ± 3.16	84.01 ± 2.08	79.04 ± 4.08
IEDG-BERT	85.99 ± 1.08	87.57 ± 0.30	88.34 ± 3.64	86.12 ± 1.85
IED-RoBERTa	81.47 ± 3.08	86.91 ± 2.14	88.00 ± 0.81	87.51 ± 1.05
IEDG-RoBERTa	84.78 ± 3.00	92.41 ± 0.98	91.65 ± 4.19	88.04 ± 4.12
IED-DistilBERT	75.70 ± 0.11	80.53 ± 0.57	83.25 ± 2.05	81.85 ± 1.92
IEDG-DistilBERT	80.63 ± 0.33	83.24 ± 4.94	84.20 ± 2.68	85.32 ± 2.36
IED-ALBERT	79.88 ± 5.23	80.74 ± 2.25	81.05 ± 1.23	84.11 ± 1.03
IEDG-ALBERT	82.99 ± 2.93	81.46 ± 0.50	82.74 ± 0.25	85.16 ± 0.29

Table 4: Ablation Study of IED/IEGD with Different Classifier Backbones on Benchmark Datasets.

Gain provides more discriminative features for distinguishing between human-written and LLM-generated text than using Information Entropy alone. The IEGD feature engineering approach is more effective than the IED. Using more powerful classifier backbones like RoBERTa and BERT with IEGD features yields the best performance. Furthermore, even lighter models like DistilBERT can achieve competitive results when combined with the potent IEGD features. This demonstrates that the robustness of our feature engineering can compensate for reduced model complexity, making them viable options when computational resources are a concern.

5 Conclusion

We proposed a model-agnostic framework for detecting LLM-generated text named IEGD. Our findings confirm that human-written content exhibits inherently higher information entropy than the structured, predictable patterns produced by LLMs. IEGD leverages this distinction by quantifying uncertainty reduction, transforming subtle statistical variations into discriminative signals. Future work will incorporate conditional entropy to capture fine-grained distributional dynamics. By offering token-level resolution, this approach aims to overcome the limitations of aggregate metrics like perplexity, exposing the specific linguistic pacing essential for detecting high-fluency generated content.

Limitations

The empirical validation of our methods predominantly relies on datasets such as XSum, SQUAD, TOEFL, and Reddit WritingPrompts, which are primarily English-based. Future research needs to explore the information entropy characteristics of different languages, potentially revealing unique patterns in entropy distribution for certain languages (such as Chinese) due to the characteristics of their writing systems. Furthermore, the potential for adversarial attacks specifically designed to circumvent entropy-based detection (e.g., by manipulating text to match expected human entropy profiles) poses an ongoing challenge [Liu *et al.*, 2026a]. While our methods provide a strong baseline, their long-term robustness against these evolving threats will require continuous evaluation and potential adaptation, possibly by incorporating features resilient to such manipulations or developing dynamic thresholding mechanisms [Huo *et al.*, 2025].

Acknowledgments

This work is supported by the National Key Research and Development Program of China under grant 2024YFC3307900; the National Natural Science Foundation of China under grants 62302184, 62376103, 62436003 and 62206102; Major Science and Technology Project of Hubei Province under grant 2024BAA008; Hubei Science and Technology Talent

Service Project under grant 2024DJC078; Ant Group through CCF-Ant Research Fund. The computation is completed in the HPC Platform of Huazhong University of Science and Technology. Hong Kong RGC General Research Fund (Grant Nos. 15221123, 15216424, and 15211525), and the Hong Kong PolyU Internal Research Fund (Grant Nos. P0058468 and P0056171).

References

[Antoun *et al.*, 2023] Wissam Antoun, Benoît Sagot, and Djamé Seddah. From text to source: Results in detecting large language model-generated content. *arXiv preprint arXiv:2309.13322*, 2023.

[Bakhtin *et al.*, 2019] Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*, 2019.

[Bao *et al.*, 2023] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*, 2023.

[Bhattacharjee *et al.*, 2024] Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. Eagle: A domain generalization framework for ai-generated text detection. *arXiv preprint arXiv:2403.15690*, 2024.

[Black *et al.*, 2022] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dugan *et al.*, 2023] Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12763–12771, 2023.

[Fan *et al.*, 2018] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

[Fröhling and Zubiaga, 2021] Leon Fröhling and Arkaitz Zubiaga. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Computer Science*, 7:e443, 2021.

[Gao *et al.*, 2018] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE, 2018.

[Gao *et al.*, 2020] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

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

[Huo *et al.*, 2025] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Kirchenbauer *et al.*, 2023] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023.

[Lancaster, 2023] Thomas Lancaster. Artificial intelligence, text generation tools and chatgpt–does digital watermarking offer a solution? *International Journal for Educational Integrity*, 19(1):10, 2023.

[Li *et al.*, 2024a] Shiwei Li, Zhuoqi Hu, Xing Tang, Haozhao Wang, Shijie Xu, Weihong Luo, Yuhua Li, Xiuqiang He, and Ruixuan Li. Mixed-precision embeddings for large-scale recommendation models. *arXiv preprint arXiv:2409.20305*, 2024.

[Li *et al.*, 2024b] Shiwei Li, Wenchao Xu, Haozhao Wang, Xing Tang, Yining Qi, Shijie Xu, Weihong Luo, Yuhua Li, Xiuqiang He, and Ruixuan Li. Fedbat: communication-efficient federated learning via learnable binarization. In *Proceedings of the 41st International Conference on Machine Learning*, pages 29074–29095, 2024.

[Li *et al.*, 2025a] Yichen Li, Haozhao Wang, Yining Qi, Wei Liu, and Ruixuan Li. Re-fed+: A better replay strategy for federated incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[Li *et al.*, 2025b] Yichen Li, Haozhao Wang, Wenchao Xu, Tianzhe Xiao, Hong Liu, Minzhu Tu, Yuying Wang, Xin Yang, Rui Zhang, Shui Yu, et al. Unleashing the power of continual learning on non-centralized devices: A survey. *IEEE Communications Surveys & Tutorials*, 2025.

[Liu *et al.*, 2022] Wei Liu, Haozhao Wang, Jun Wang, Ruixuan Li, Chao Yue, and YuanKai Zhang. Fr: Folded rationalization with a unified encoder. *Advances in Neural Information Processing Systems*, 35:6954–6966, 2022.

[Liu *et al.*, 2023a] Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. Coco: Coherence-enhanced machine-generated text detection under low resource with contrastive learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16167–16188, 2023.

[Liu *et al.*, 2023b] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017, 2023.

[Liu *et al.*, 2026a] Wei Liu, Hongkai Liu, Zhiying Deng, Yee Whye Teh, and Wee Sun Lee. From backward spreading to forward replay: Revisiting target construction in llm parameter editing. *arXiv preprint arXiv:2605.00358*, 2026.

[Liu *et al.*, 2026b] Wei Liu, Haomei Xu, Hongkai Liu, Zhiying Deng, Ruixuan Li, Heng Huang, Yee Whye Teh, and Wee Sun Lee. Are we evaluating the edit locality of llm model editing properly? *arXiv preprint arXiv:2601.17343*, 2026.

[Mitchell *et al.*, 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.

[Narayan *et al.*, 2018] Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.

[Quidwai *et al.*, 2023] Mujahid Ali Quidwai, Chunhui Li, and Parijat Dube. Beyond black box ai-generated plagiarism detection: From sentence to document level. *arXiv preprint arXiv:2306.08122*, 2023.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Rajpurkar *et al.*, 2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[Sadasivan *et al.*, 2023] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

[Shi *et al.*, 2024] Yuhui Shi, Qiang Sheng, Juan Cao, Hao Mi, Beizhe Hu, and Danding Wang. Ten words only still help: Improving black-box ai-generated text detection via proxy-guided efficient re-sampling. *arXiv preprint arXiv:2402.09199*, 2024.

[Taguchi *et al.*, 2024] Kaito Taguchi, Yujie Gu, and Kouichi Sakurai. The impact of prompts on zero-shot detection of ai-generated text. *arXiv preprint arXiv:2403.20127*, 2024.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Wang *et al.*, 2023a] Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. SeqXGPT: Sentence-level AI-generated text detection. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1144–1156, Singapore, December 2023. Association for Computational Linguistics.

[Wang *et al.*, 2023b] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, et al. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*, 2023.

[Wang *et al.*, 2024] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[Wang *et al.*, 2025a] Bingbing Wang, Zhengda Jin, Bin Liang, Jing Li, and Ruifeng Xu. Remod: Rethinking modality contribution in multimodal stance detection via dual reasoning. *arXiv preprint arXiv:2511.06057*, 2025.

[Wang *et al.*, 2025b] Bingbing Wang, Jingjie Lin, Zhixin Bai, Zihan Wang, Shengzhe Sun, Zhengda Jin, Zhiyuan Wen, Geng Tu, Jing Li, Erik Cambria, et al. Internet meme on social media: A comprehensive review and new perspectives. *Information Fusion*, page 104102, 2025.

[Wang *et al.*, 2026] Bingbing Wang, Jing Li, and Ruifeng Xu. Bounded state in an infinite horizon: Proactive hierarchical memory for ad-hoc recall over streaming dialogues. *arXiv preprint arXiv:2603.04885*, 2026.

[Wu *et al.*, 2023] Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. Llmdet: A third party large language models generated text detection tool. *arXiv preprint arXiv:2305.15004*, 2023.

[Wu *et al.*, 2024] Junchao Wu, Runzhe Zhan, Derek F Wong, Shu Yang, Xuebo Liu, Lidia S Chao, and Min Zhang. Who wrote this? the key to zero-shot llm-generated text detection is gecscore. *arXiv preprint arXiv:2405.04286*, 2024.

[Zhang *et al.*, 2022] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.