

592	Contents	
593	1 Introduction	1
594	2 Related Works	3
595	3 Methods	4
596	3.1 Problem Reformulation: Why Binary Classification Fails for Human Text Detection?	4
597	3.2 Method Overview: OOD Detection Framework for Machine-Generated Text	5
598	3.3 Alternative OOD Detection Losses: HRN and Energy-Based Methods	6
599	3.4 Overall Training Objectives	7
600	4 Experiments	7
601	4.1 Experimental Setup	7
602	4.2 Main Results	7
603	4.3 Experimental Analysis	8
604	5 Conclusion	9
605	A Proof of Theoretical Results	16
606	A.1 Preliminaries	16
607	A.2 Analysis of the Pearson χ^2 Divergence	16
608	A.3 Proof of Theorem 2	17
609	A.4 Alternative View: Binary Classifiers Overfit to Slightly Defected Datasets	17
610	B Detail of Experimental Setup	19
611	B.1 Datasets	19
612	B.2 Baseline Setup	19
613	B.3 Implementation Detail	20
614	C More Experimental Results	20
615	C.1 Hyper-parameter Sensitivity Analysis	20
616	C.2 More Metrics	21
617	C.3 Different Backbones	21
618	C.4 Limitation and Ethical Statement	21

A Proof of Theoretical Results

In this section, we provide formal justification for Theorem 2 discuss the magnitude of Pearson χ^2 divergence, and provide an alternative theorem to further characterize the inevitability of overfitting binary classifiers to biased human-written text distributions.

A.1 Preliminaries

We use integral $\int_{x \in \mathcal{X}} \cdot dx$ to equivalently denote integral or summation over the data space \mathcal{X} . For the ground truth classifier $\hat{p}_M(x)$, we introduce the following auxiliary definitions to assist the writing:

$$\begin{aligned}\hat{p}_H(x) &:= \hat{p}_M(x) \\ \hat{p}(y=0|x) &:= \hat{p}_H(x), \quad \hat{p}(y=1|x) := \hat{p}_M(x)\end{aligned}$$

Before proving the theorem, we characterize the cross-entropy loss \mathcal{L}_{CE} :

$$\begin{aligned}\mathcal{L}_{\text{CE}}(f_\theta|\mathcal{D}) &= -q_M \mathbb{E}_{x \sim P_M} [\log p_\theta(y=0|x)] - q_H \mathbb{E}_{x \sim P_H} [\log p_\theta(y=1|x)] \\ &= - \int_{x \in \mathcal{X}} \left[q_M P_M(x) \log p_\theta(y=0|x) + q_H P_H(x) \log p_\theta(y=1|x) \right] dx \\ &= - \int_{x \in \mathcal{X}} \left[\hat{p}_M(x) \log p_\theta(y=0|x) + (1 - \hat{p}_M(x)) \log p_\theta(y=1|x) \right] P_{\mathcal{D}}(x) dx \\ &= \int_{x \in \mathcal{X}} \mathcal{H}(\hat{p}(\cdot|x), p_\theta(\cdot|x)) P_{\mathcal{D}}(x) dx,\end{aligned}\tag{11}$$

where the third equality comes from Assumption 1, $P_{\mathcal{D}}(x) := q_M P_M(x) + q_H P_H(x)$ denotes the joint probability at x of dataset \mathcal{D} . and $\mathcal{H}(\cdot, \cdot)$ denotes the cross-entropy between two distributions. From the properties of cross-entropy, this is minimized when $\hat{p} = p_\theta$, i.e. when the model predictions exactly match the ground truth, with minimal value

$$\min_{\theta} \mathcal{L}_{\text{CE}}(f_\theta|\mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}} \mathcal{H}(\hat{p}(\cdot|x)).\tag{12}$$

We also define The Pearson χ^2 divergence below for future reference.

Definition 3. For two distributions P_1, P_2 over set \mathcal{X} , the Pearson χ^2 divergence between P_1 and P_2 is defined as

$$D_{\chi^2}(P_1\|P_2) = \int_{x \in \mathcal{X}} \frac{(P_1(x) - P_2(x))^2}{P_2(x)} dx = \int_{x \in \mathcal{X}} \frac{P_1^2(x)}{P_2(x)} dx - 1\tag{13}$$

A.2 Analysis of the Pearson χ^2 Divergence

The effectiveness of Theorem 2 depends on the magnitude of the Pearson χ^2 divergence $D_{\chi^2}(\hat{P}_H\|P_H)$: the larger this divergence, the larger the cross-entropy loss for a linear classifier well-fitted to the training set. Here we relate this value to the intuition that the training human-text data distribution is biased and cannot reflect open-world distribution.

We model the intuition of “open-world” distribution \hat{P}_H by considering a training distribution P_H that is a **shifted biased distribution** of \hat{P}_H . More concretely, this means that there is a subset \mathcal{X}_0 of \mathcal{X} , such that

$$P_H(x) = \begin{cases} C_1 \hat{P}_H(x), & x \in \mathcal{X}_0, \\ C_2 \hat{P}_H(x), & x \notin \mathcal{X}_0, \end{cases}$$

where C_1, C_2 are constants such that $C_1 \mu_{\hat{P}_H}(\mathcal{X}_0) + C_2 [1 - \mu_{\hat{P}_H}(\mathcal{X}_0)] = 1$ and $C_1 \gg C_2$, and $\mu_{\hat{P}_H}$ denotes the measure with respect to \hat{P}_H . Such a distribution follows the same probability ratios as \hat{P}_H within and without \mathcal{X}_0 respectively, but is very close to 0 outside \mathcal{X}_0 , hence most of the samples come from the partial region \mathcal{X}_0 . This is of course only an approximation of a realistic scenario, but

644 it captures the essence of training data being only a partial observation of the whole picture. With
 645 this, the Pearson χ^2 divergence is

$$\begin{aligned} D_{\chi^2}(\hat{P}_H \| P_H) &= \int_{x \in \mathcal{X}} \frac{\hat{P}_H^2(x)}{P_H(x)} dx - 1 \\ &= \int_{x \in \mathcal{X}_0} \frac{1}{C_1} \hat{P}_H(x) dx + \int_{x \in \mathcal{X} \setminus \mathcal{X}_0} \frac{1}{C_2} \hat{P}_H(x) dx - 1 \\ &= \frac{\mu_{\hat{P}_H}(\mathcal{X}_0)}{C_1} + \frac{1 - \mu_{\hat{P}_H}(\mathcal{X}_0)}{C_2} - 1, \end{aligned}$$

646 which **blows up to infinity** inverse-linearly as $C_2 \rightarrow 0$, corresponding to the likely case where the
 647 dataset is heavily biased towards a partial region of the text data space.

648 A.3 Proof of Theorem 2

649 *Proof of Theorem 2* Let $\Delta(x) = \mathcal{H}(\hat{p}(\cdot|x), p_\theta(\cdot|x)) - \mathcal{H}(\hat{p}(\cdot|x))$ be the suboptimality of f_θ at x .
 650 From (11),

$$\mathcal{L}_{\text{CE}}(f_\theta | \mathcal{D}) - \mathbb{E}_{x \sim \mathcal{D}} \mathcal{H}(\hat{p}(\cdot|x)) = \int_{x \in \mathcal{X}} \Delta(x) P_{\mathcal{D}}(x) dx, \quad (14)$$

$$\mathcal{L}_{\text{CE}}(f_\theta | \hat{\mathcal{D}}) - \mathbb{E}_{x \sim \hat{\mathcal{D}}} \mathcal{H}(\hat{p}(\cdot|x)) = \int_{x \in \mathcal{X}} \Delta(x) P_{\hat{\mathcal{D}}}(x) dx, \quad (15)$$

651 Since cross-entropy is unbounded for $p_\theta(\cdot|x)$ given any $\hat{p}(\cdot|x)$, we can choose a classifier f_θ such
 652 that $\Delta(x) = \Delta_0 \times P_{\hat{\mathcal{D}}}(x)/P_{\mathcal{D}}(x)$ for any $x \in \mathcal{X}$ given a target suboptimality Δ_0 . With this we have
 653 the training suboptimality

$$\int_{x \in \mathcal{X}} \Delta(x) P_{\mathcal{D}}(x) dx = \int_{x \in \mathcal{X}} \Delta_0 P_{\hat{\mathcal{D}}}(x) dx = \Delta_0,$$

654 while the loss on ground truth dataset $\hat{\mathcal{D}}$ is

$$\begin{aligned} \int_{x \in \mathcal{X}} \Delta(x) P_{\hat{\mathcal{D}}}(x) dx &= \int_{x \in \mathcal{X}} \Delta_0 \times \frac{P_{\hat{\mathcal{D}}}^2(x)}{P_{\mathcal{D}}(x)} dx \\ &= \Delta_0 \int_{x \in \mathcal{X}} \frac{(q_M \hat{P}_M(x) + q_H \hat{P}_H(x))^2}{q_M P_M(x) + q_H P_H(x)} dx. \end{aligned} \quad (16)$$

655 Now notice that from Assumption 1, the ratios

$$\frac{q_M \hat{P}_M(x)}{q_H \hat{P}_H(x)} = \frac{q_M P_M(x)}{q_H P_H(x)} = \frac{\hat{p}_M(x)}{\hat{p}_H(x)},$$

656 therefore we can write (16) as

$$\begin{aligned} \int_{x \in \mathcal{X}} \Delta(x) P_{\hat{\mathcal{D}}}(x) dx &= \Delta_0 \int_{x \in \mathcal{X}} \left(\frac{\hat{p}_M(x)}{\hat{p}_H(x)} + 1 \right) \cdot q_H \frac{\hat{P}_H^2(x)}{P_H(x)} dx \\ &\geq \Delta_0 \int_{x \in \mathcal{X}} q_H \frac{\hat{P}_H^2(x)}{P_H(x)} dx \\ &= q_H \Delta_0 [D_{\chi^2}(\hat{P}_H \| P_H) + 1], \end{aligned}$$

657 thus finishing the proof. □

658 A.4 Alternative View: Binary Classifiers Overfit to Slightly Defected Datasets

659 While theoretically reasonable, Assumption 1 usually does not hold for many data distributions. For
 660 example, if a training human-text distribution consists only of one “style” of texts, then the probability
 661 that some human-text x^* of another style is sampled from this distribution would be close to 0, even
 662 smaller than the probability that x^* is sampled from the machine process, i.e. $P_H(x^*) \ll P_M(x^*)$.

This leads to a violation of Assumption [1](#) in that the Bayes probability that x^* is human-generated does not match $\hat{p}_M(x)$. In this scenario, binary classifiers actually can overfit to the training dataset and fail to generalize by a nonzero gap.

To illustrate this, we characterize this dataset defect in the following definition in place of Assumption [1](#).

Definition 4. For a machine-human text data distribution $\mathcal{D} = \{q_M, q_H, P_M, P_H\}$, define its **posterior binary classifier** to be

$$p_{\mathcal{D}}(y = 1|x) = \frac{q_M P_M(x)}{q_M P_M(x) + q_H P_H(x)},$$

$$p_{\mathcal{D}}(y = 0|x) = \frac{q_H P_H(x)}{q_M P_M(x) + q_H P_H(x)}.$$

We define the **kwality** of \mathcal{D} as

$$\mathbf{K}(\mathcal{D}) := \mathbb{E}_{x \sim P_{\mathcal{D}}(x)} D_{\text{KL}}[\hat{p}(\cdot|x) \| p_{\mathcal{D}}(\cdot|x)],$$

which measures the expected KL-divergence between $p_{\mathcal{D}}$ and \hat{p} over joint text distribution $P_{\mathcal{D}}(x) = q_M P_M(x) + q_H P_H(x)$, and say dataset \mathcal{D} is δ -**suboptimal** if its kwality $\mathbf{K}(\mathcal{D}) < \delta$.

Remark 5. A few things are worthy of note:

- When the posterior distribution $p_{\mathcal{D}} \rightarrow \hat{p}$, kwality reaches its minimum value 0, while higher kwality values correspond to deviations in labeling and hence degradation of dataset.
- The idea behind this definition is that, realistically we cannot expect the training dataset to be perfectly accurate, and kwality more or less measures the deviation of the model from the ground truth classifier.
- The expectation in kwality is taken over text distribution $P(x)$, which means mislabeling data in *sparser* text regions are less detrimental to kwality than *denser* regions. This aligns with the fact that sparse regions contain less data points within the sampled dataset, where accuracy cannot be guaranteed, but also warrants less attention from the learning model.

Our main result is that fully training models on a near-optimal dataset can lead to overfitting, and result in a large validation loss on open-world data distribution.

Theorem 6. There exists a δ -suboptimal dataset \mathcal{D} such that for any binary classifier that achieves optimal cross-entropy loss on \mathcal{D} , the generalization loss on open-world dataset $\widehat{\mathcal{D}}$ which complies with \hat{p} (Assumption [1](#)) is at least $\delta D_{\chi^2}(P_{\mathcal{D}} \| P_{\widehat{\mathcal{D}}})$.

Proof. From optimal cross-entropy loss, we get from [\(11\)](#) that

$$\begin{aligned} 0 &= \mathcal{L}_{\text{CE}}(f_{\theta} | \mathcal{D}) - \mathbb{E}_{x \sim \mathcal{D}} \mathcal{H}(p_{\mathcal{D}}(\cdot|x)) \\ &= \int_{x \in \mathcal{X}} \left[\mathcal{H}(p_{\mathcal{D}}(\cdot|x), p_{\theta}(\cdot|x)) - \mathcal{H}(p_{\mathcal{D}}(\cdot|x)) \right] P_{\mathcal{D}}(x) dx, \end{aligned}$$

which means for any $x \in \mathcal{X}$,

$$\mathcal{H}(p_{\mathcal{D}}(\cdot|x), p_{\theta}(\cdot|x)) = \mathcal{H}(p_{\mathcal{D}}(\cdot|x)) \Leftrightarrow p_{\theta}(\cdot|x) = p_{\mathcal{D}}(\cdot|x).$$

Now consider the generalization loss:

$$\begin{aligned} &\mathcal{L}_{\text{CE}}(f_{\theta} | \widehat{\mathcal{D}}) - \mathbb{E}_{x \sim \widehat{\mathcal{D}}} \mathcal{H}(\hat{p}(\cdot|x)) \\ &= \int_{x \in \mathcal{X}} \left[\mathcal{H}(\hat{p}(\cdot|x), p_{\theta}(\cdot|x)) - \mathcal{H}(\hat{p}(\cdot|x)) \right] P_{\widehat{\mathcal{D}}}(x) dx \\ &= \int_{x \in \mathcal{X}} D_{\text{KL}}[\hat{p}(\cdot|x) \| p_{\mathcal{D}}(\cdot|x)] P_{\widehat{\mathcal{D}}}(x) dx, \end{aligned}$$

Since KL-divergence is unbounded, we can choose $\hat{p}(\cdot|x)$ for each $x \in \mathcal{X}$ such that $D_{\text{KL}}[\hat{p}(\cdot|x) \| p_{\mathcal{D}}(\cdot|x)] = \delta P_{\widehat{\mathcal{D}}}(x) / P_{\mathcal{D}}(x)$, which guarantees $\mathbf{K}(\mathcal{D}) = \delta$, while

$$\mathcal{L}_{\text{CE}}(f_{\theta} | \widehat{\mathcal{D}}) - \mathbb{E}_{x \sim \widehat{\mathcal{D}}} \mathcal{H}(\hat{p}(\cdot|x)) \geq \int_{x \in \mathcal{X}} \delta \times \frac{P_{\widehat{\mathcal{D}}}^2(x)}{P_{\mathcal{D}}(x)} dx = \delta D_{\chi^2}(P_{\mathcal{D}} \| P_{\widehat{\mathcal{D}}}).$$

□

B Detail of Experimental Setup

B.1 Datasets

In this section, we discuss more details of our dataset in our experiments.

- **DeepFake.** The Deepfake dataset comprises text generated by 27 large language models (LLMs) alongside human-written content sourced from multiple websites across 10 distinct domains, totaling 332K training and 57K test samples. It defines six diverse evaluation scenarios, including cross-domain, unseen domain, and model detection setting. In our main experiment, we use the cross-domain and cross-model setting, which includes various domains and models in both the training and testing set. For the generalizability validation experiments, we utilize the Unseen Domains and Unseen Models of DeepFake, where the former setting means there are unseen-domain texts in the testing set and the latter setting indicates that there are texts generated by unseen LLMs during testing.
- **M4.** The M4 dataset is a comprehensive benchmark spanning multiple domains, models, and languages, comprising data from 8 large language models (LLMs), 6 domains, and 9 languages. It includes human-written content sourced from platforms such as Wikipedia, WikiHow [76], Reddit, arXiv, and PeerRead [77]. Leveraging human-authored prompts, models including ChatGPT [1], DaVinci-003, LLaMA [78], FLAN-T5 [79], Cohere [80] and so on generate outputs across nine languages, including English, Chinese, and Russian. As part of the M4 initiative, a competition [81] is organized to evaluate the detection of AI-generated text at both the paragraph and sentence levels. Two evaluation settings are used: monolingual and multilingual. The multilingual setting introduces new languages absent from the training and validation sets, with AI-generated texts also undergoing paraphrasing. We focus on multilingual setting since this is a more complicated and hard setting for correct LLM detection. The multilingual settings includes 157K training and 42K testing data. The test dataset of M4 also includes the unseen models from the training set.
- **RAID.** RAID [73] is the largest and most comprehensive dataset available for evaluating AI-generated text detection systems. It comprises over 10 million documents, encompassing 11 large language models (LLMs), 11 content genres, 4 decoding strategies, and 12 types of adversarial attacks. The detailed information of RAID dataset can be found in Table 4. We hold out 10% of training data as validation set for evaluation, which includes the texts with all kinds of attacks. We utilize the RAID dataset to compare the robustness of different methods on the attacked texts. During training, we filter the attacked samples in the training set to train the model.

Category	Values
Models	ChatGPT, GPT-4, GPT-3 (text-davinci-003), GPT-2 XL, Llama 2 70B (Chat), Cohere, Cohere (Chat), MPT-30B, MPT-30B (Chat), Mistral 7B, Mistral 7B (Chat)
Domains	ArXiv Abstracts, Recipes, Reddit Posts, Book Summaries, NYT News Articles, Poetry, IMDb Movie Reviews, Wikipedia, Czech News, German News, Python Code
Decoding Strategies	Greedy (T=0), Sampling (T=1), Greedy + Repetition Penalty (T=0, $\theta=1.2$), Sampling + Repetition Penalty (T=1, $\theta=1.2$)
Adversarial Attacks	Article Deletion, Homoglyph, Number Swap, Paraphrase, Synonym Swap, Misspelling, Whitespace Addition, Upper-Lower Swap, Zero-Width Space, Insert Paragraphs, Alternative Spelling

Table 4: RAID dataset composition. The table includes the information about the models, domains, decoding strategies, and adversarial attack kinds.

B.2 Baseline Setup

In this section, we discuss the details of the experiments setup on baseline models including zero-shot and training-based baselines. For DetectLLM, DetectGPT, DNA-GPT and FastDetectGPT, we

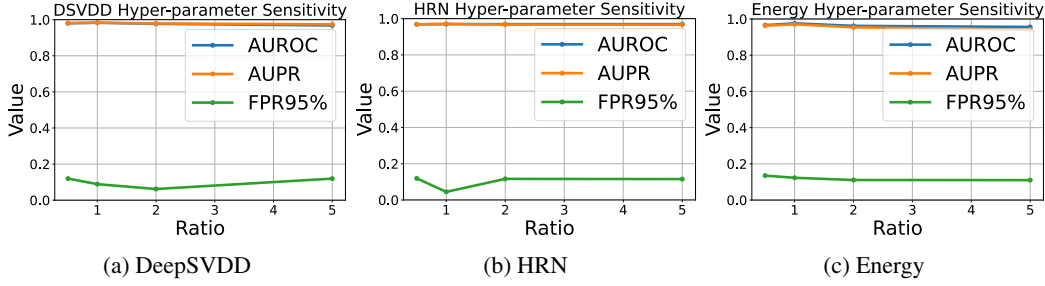


Figure 3: Hyper-parameter sensitivity analysis of our method. The experiments are conducted on DeepFake dataset and show that our method is robust to the choice of weight.

utilize the official implementation of FastDectGPT¹ which also includes the implementation of other methods. We utilize GPT-Neo-2.7B as the scoring model for all methods. T5-Small is used as the sampling model to do perturbation for DetectGPT and DetectLLM. We do 100 perturbation for each text. Due to the low speed of perturbation, we randomly select 10K samples from each benchmark for evaluation for DetectLLM, DetectGPT and DNA-GPT. For DALD, we use the LoRA model trained by GPT-4 texts (which is demonstrated with the best generalizability in the original paper) based on Llama-2-7B as the scoring model. We follow the same experimental setting of Binoculars and apply the Falcon model to compute the final metric. For Glimpse, we utilize the official implementation and call OpenAI davinci-002 API to compute the geometric metric.

For training-based method, such as GPT-Zero, we utilize the open-source version for evaluation². For RADAR and GhostBuster, we apply their official implementation for testing with the updated OpenAI API model. Moreover, we adopt the official open-source implementation of BiScope and utilize gemma-2b as the scoring model.

B.3 Implementation Detail

We adopt three OOD detection methods including DeepSVDD, HRN and Energy-based methods for our experiments. For DeepSVDD, we use the machine texts from the training set to compute the initial center and compute the corresponding loss with the center. We freeze the parameters of the center point and disable the optimization on the center point. During inference, we compute the L_2 distance of the given sample and the center point as the probability of being human-written text. For HRN, follow its original setting by training a one-class classifier per model family using its corresponding data as the positive class and averaging their scores at inference time. We also follow the original hyperparameter settings, where $\lambda = 0.1$ and $n = 12$. For energy-based method, a classification head is attached following the backbone model. We follow [18] to choose hyper-parameters, where $m_{in} = -27$ and $m_{out} = -5$. DeepSVDD is trained from scratch, and HRN and Energy load the pre-trained weights of DeTeTive. The learning-rate is set as $2e-5$ with batch size 32 per device (64 global batch size in our experiments), and the optimizer is Adam [75] with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The loss weights α and β are set as 1 in our experiments. We train all model for 20 epochs. All experiments are conducted on 2*A100 GPUs.

C More Experimental Results

C.1 Hyper-parameter Sensitivity Analysis

For the hyperparameter introduced by the DeepSVDD, HRN and Energy-based method, we use the default setting from their original paper. In our work, the hyperparameter we introduce is the ratio of contrastive loss and OOD detection loss, namely $\alpha : \beta$. This ratio balances the contrastive loss and the OOD loss. We conduct a set of experiments with different the loss ratio settings on DeepFake

¹<https://github.com/baoguangsheng/fast-detect-gpt>

²<https://github.com/BurhanUITayyab/GPTZero>

Table 5: Performance comparison with the baseline on more evaluation metrics including accuracy and F1 score.

Method	RAID	
	Accuracy \uparrow	F1 \uparrow
DeTeCtive	96.5	55.2
Ours (DSVDD)	98.6	70.0
Ours (HRN)	98.5	67.7
Ours (Energy)	98.7	99.4

Table 6: Different backbone model.

Backbone Model	AUROC \uparrow	AUPR \uparrow	FPR95 \downarrow
BERT _{base}	96.9	96.7	13.0
BERT _{large}	93.1	94.1	45.3
RoBERTa _{base}	96.1	96.5	16.1
FLAN-T5 _{base}	95.3	95.6	18.9
FLAN-T5 _{large}	96.9	96.5	9.7
SimCSE-BERT _{base}	96.8	96.9	17.7
SimCSE-RoBERTa _{base}	98.3	98.3	8.9

dataset and the results are shown Figure 3. Overall, the results are stable and robust to the choice of weight. Ratio 1 : 1 is a great choice which balanced well for all methods and metrics.

C.2 More Metrics

Besides the AUROC, AUPR and FPR at TPR 95%, we also provide other metrics including the accuracy and F1 score for the comparison. We conduct the experiments on RAID dataset and report the results in Table 5. All settings of our method achieve better performance than the baseline on both accuracy and F1. Our energy-based method shows significant improvement, obtaining 98.7 and 99.4 on accuracy and F1, respectively.

C.3 Different Backbones

To show the affection of the backbone model for the LLM detection performance, we conduct a further experimental analysis with different training backbones on DeepFake dataset, as shown in Table 6. We can observe that our method is robust to the choice of model backbone. All models generate reasonable results which high AUROC, AUPR and low FRP values.

C.4 Limitation and Ethical Statement

In this paper, we theoretically and empirically analyze why and how to model LLM text detection task as the out-of-distribution detection task. However, current detection can only tell whether the given text is AI-generated or human-written, which lacks further interpretation about how the text is generated by LLM. Therefore, we hope the user uses the detection result as a reference and doesn't make decisions solely based on the detection results, especially in areas such as academia.