Reliably Bounding False Positives: A Zero-Shot Machine-Generated Text Detection Framework via Multiscaled Conformal Prediction

Xiaowei Zhu[1,2], Yubing Ren[1,2*], Yanan Cao[1,2], Xixun Lin[1,2], Fang Fang[1,2], Yangxi Li[3]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]National Computer Network Emergency Response Technical Team, Coordination Center of China, Beijing, China
{zhuxiaowei, renyubing}@iie.ac.cn

Abstract

The rapid advancement of large language models has raised significant concerns regarding their potential misuse by malicious actors. As a result, developing effective detectors to mitigate these risks has become a critical priority. However, most existing detection methods focus excessively on detection accuracy, often neglecting the societal risks posed by high false positive rates (FPRs). This paper addresses this issue by leveraging Conformal Prediction (CP), which effectively constrains the upper bound of FPRs. While directly applying CP constraints FPRs, it also leads to a significant reduction in detection performance. To overcome this trade-off, this paper proposes a Zero-Shot Machine-Generated Text Detection Framework via Multiscaled Conformal Prediction (MCP), which both enforces the FPR constraint and improves detection performance. This paper also introduces RealDet, a high-quality dataset that spans a wide range of domains, ensuring realistic calibration and enabling superior detection performance when combined with MCP. Empirical evaluations demonstrate that MCP effectively constrains FPRs, significantly enhances detection performance, and increases robustness against adversarial attacks across multiple detectors and datasets. Dataset and code are available at https://github.com/Xiaoweizhu57/RealDet.

1 Introduction

The rapid advancement of large language models (LLMs) has led to the generation of fluent, natural, and high-quality text that increasingly resembles human-written text. LLMs are being leveraged to enhance productivity across various domains, including news reporting, storytelling, and academic research (Alshater, 2022; Yuan et al., 2022; Christian, 2023), significantly contributing to both industrial and academic progress. However, this same capability has also opened the door for misuse, with malicious actors exploiting LLMs to generate fake news (Ahmed et al., 2021), spam (Guo et al., 2021), malicious reviews (Adelani et al., 2019), and other harmful contents that pose substantial risks to society. As a result, developing advanced Machine-Generated Text (MGT) detectors has become an urgent necessity.

Researchers have proposed numerous methods for MGT detection, including zero-shot detectors based on statistical metrics (Bao et al., 2024; Hans et al., 2024; Mitchell et al., 2023) and supervised detectors fine-tuned on pretrained models (Solaiman et al., 2019b; Conneau et al., 2019). However, these approaches excessively emphasize detection accuracy while neglecting the potential societal harm caused by high false positive rates (FPRs). This concern is consistent with the findings of Dugan et al. (2024), who highlighted that existing detectors often exhibit dangerously high FPRs under default thresholds. Detectors with high FPRs are impractical for real-world applications, as they fail to provide reliable guidance to users.

In this paper, we propose leveraging conformal prediction (CP) (Vovk et al., 1999) to address the challenges of high false positive rates (FPRs) in machine-generated text (MGT) detection. CP provides an upper bound on the FPR, ensuring that the detection results are reliable. While directly applying CP can constrain the FPR, it may also allow certain machine-generated texts to evade de-

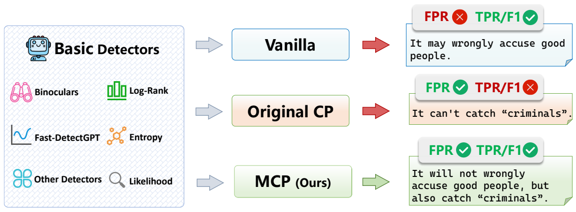

Figure 1: Detection performance of detectors under different framework configurations.

*Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12298–12319
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

tection, which would degrade overall detection performance. To address this issue, we propose a novel solution, the Zero-Shot Machine-Generated Text Detection Framework via **M**ultiscaled **C**onformal **P**rediction (**MCP**). MCP not only effectively constrains the FPR but also improves detection performance without additional training.

The MCP framework operates in the following high-level manner: First, we sample both calibration and test sets from the target datasets. Next, we select a baseline detector and compute its nonconformity scores. From the calibration set's nonconformity scores, we derive multiscaled quantiles, which act as thresholds for detection. These quantiles ensure that the FPR is constrained. Finally, we apply these thresholds to detect MGT instances in new, unseen data. Additionally, we introduce **RealDet**, a high-quality benchmark dataset designed to simulate realistic scenarios for MGT detection. RealDet is essential for ensuring that the calibration set reflects the true distribution of human-written text (HWT), addressing the gap in existing datasets.

Extensive experiments consistently demonstrate that the MCP framework effectively constrains the upper bound of the FPR while simultaneously improving detection performance. In adversarial scenarios, the MCP significantly enhances robustness. Our contributions are summarized as:

- We are the first to introduce CP into MGT detection and provide an in-depth exploration of potential optimization mechanisms.
- We propose MCP, a zero-shot detection framework that not only constrains the FPR upper bound but also improves detection performance and enhances robustness against adversarial attacks.
- We construct RealDet, the large-scale and comprehensive bilingual benchmark, consisting of 836k raw texts spanning 15 representative domains, 22 popular and powerful LLMs, and covering two adversarial attacks.

2 Preliminary

Conformal Prediction. Conformal prediction (Vovk et al., 2005; Papadopoulos et al., 2002; Lei and Wasserman, 2014) is a statistical learning framework that generates reliable predictions without training. It provides statistical guarantees for the coverage of the ground truth, assuming only data exchangeability. The workflow is:

1. Split the data into a calibration set D_{cal} and a test set D_{test}, with D_{cal} containing n instances.
2. Given a model taking input x and producing output y. Then define a nonconformity score $s(x, y) \in \mathbb{R}$, where larger scores encode worse agreement between x and y.
3. Compute quantile $\hat{q}$ of s derived from D_{cal}:

$$\hat{q} = \text{quantile}\left(s(x_1, y_1), \ldots, s(x_n, y_n); \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right). \quad (1)$$

4. Using $\hat{q}$ as the prediction threshold to predict each test instance:

$$\mathcal{C}\left(X_{\text{test}}\right) = \left\{y : s\left(X_{\text{test}}, y\right) \leq \hat{q}\right\}. \quad (2)$$

Theorem 1. ***Conformal coverage guarantee*** *(Vovk et al., 1999). Suppose the calibration set $(X_i, Y_i)_{i=1,\ldots,n}$ and the new instance (X_{test}, Y_{test}) are independent and identically distributed (i.i.d.). Then, the following holds:*

$$P\left(Y_{test} \in \mathcal{C}\left(X_{test}\right)\right) \geq 1 - \alpha. \quad (3)$$

MGT Detection within CP. Given n human-written texts $(X_1, X_2, \ldots, X_n)$ as a calibration set, we are tasked to predict a new instance X_{test} is human-written or machine-generated. Based on the output of detector Det, we define a nonconformity score $s \in [0, 1]$, where a larger score indicates a lower probability that the text is human-written. Then we compute quantile $\hat{q}$ according to Equation 1 and use $\hat{q}$ as the threshold to make prediction:

$$\mathcal{C}(X_{test}) = \begin{cases} \text{HWT}, & s(X_{test}, y) \leq \hat{q} \\ \text{MGT}, & s(X_{test}, y) > \hat{q}. \end{cases} \quad (4)$$

This gives the guarantee in Theorem 1 that no more than α fraction of future human-written texts will be misclassified as machine-generated, i.e., FPR$\leq \alpha$.

3 Multiscaled Conformal Prediction

Figure 2 illustrates the MCP prediction process. First, we sample calibration and test sets from the target datasets. Next, we determine a basic detector and define its nonconformity scores. Subsequently, we derive multiscaled quantiles from the calibration set's nonconformity scores. Finally, we apply the multiscaled quantiles as the threshold to perform MGT detection on new instances.

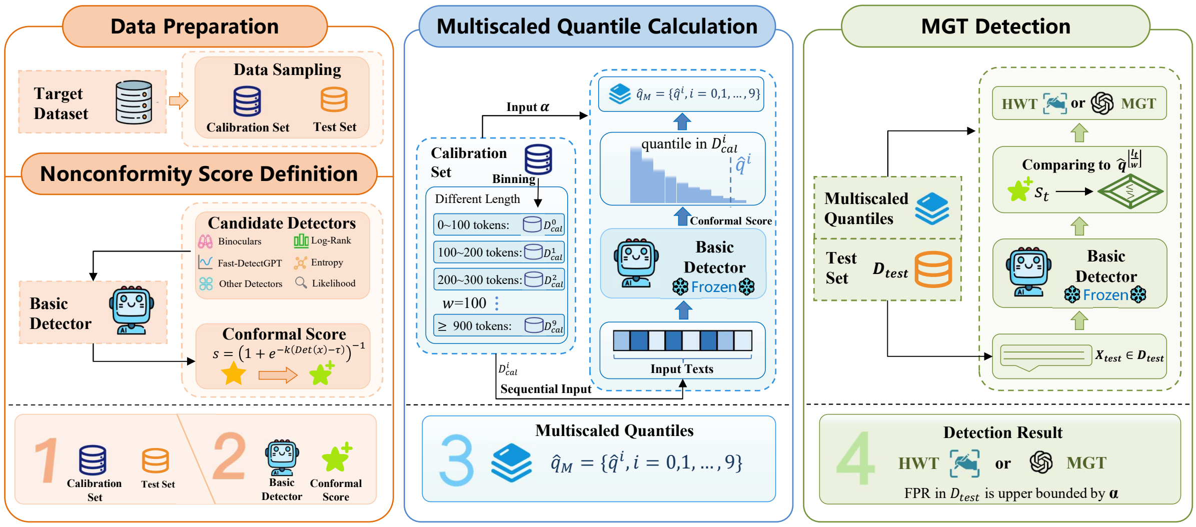

Figure 2: **The MCP Framework.** The prediction process consists of four parts, which are executed sequentially: data preparation, nonconformity score definition, multiscaled quantiles calculation, and MGT Detection.

3.1 Data Preparation

We sampled from the target dataset to create calibration and test sets, where the calibration set D_{cal} consists entirely of human-written texts, while the test set D_{test} includes both human-written and machine-generated texts. Sampling from the same dataset ensures that the human-written text in both the calibration and test sets is independent and identically distributed (i.i.d.).

3.2 Nonconformity Score Definition

First, we determine the basic detector Det, whose selection is highly flexible and can include most detectors designed for MGT detection. Then we define a nonconformity score function $s(\cdot)$ that converts the output of the basic detector $Det(x)$ into a nonconformity score s for making predictions:

$$s = (1 + e^{-k(Det(x)-\tau)})^{-1}, \tag{5}$$

where τ represents the default threshold of the basic detector, and k takes a value of either -1 or 1. A larger value of s signifies a lower probability that the input text is human-written text.

3.3 Multiscaled Quantile Calculation

Problem in traditional quantile calculation within CP. As illustrated in Figure 3, while traditional computational approaches effectively control the FPR of prediction results, they do so at the significant cost of detection performance. This trade-off prevents the detection of the majority of machine-generated texts. Based on our data analysis, we observe the following:

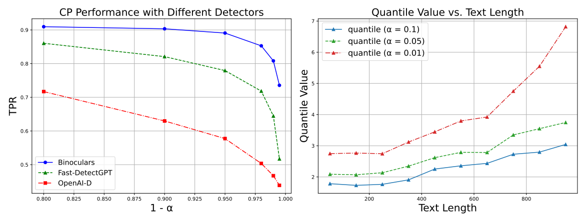

Figure 3: Left: True Positive Rate (TPR) of different detectors with the CP as a function of α. Right: Quantile values calculated for different text length intervals in Fast-DetectGPT.

Observation 1. Longer texts tend to have higher nonconformity scores. Figure 3 shows that text length significantly influences the magnitude of the quantiles. Consequently, we further calculated the Pearson correlation coefficient $\rho_{l,s}$ between text length and the nonconformity score, finding that $\rho_{l,s}$ is close to 1, which indicates a strong positive correlation. Machine-generated texts with shorter lengths and lower nonconformity scores may remain undetected, leading to a significant decline in detection performance.

Multiscaled quantiles calculation within MCP. We incorporate the positive correlation between text length and nonconformity score into the prediction process. So we perform length-aware binning on the calibration set D_{cal}, dividing it into multiple subsets $\{D^1_{cal}, D^2_{cal}, \ldots, D^K_{cal}\}$, corresponding to a specific text length interval. We employ an equal-width binning strategy, partitioning the maximum input text length $L_{\max}$ into fixed-width intervals of width w, as follows:

$$K = \lfloor \frac{L_{\max}}{w} \rfloor. \tag{6}$$

The multiscaled quantiles $\hat{q}_M$ are derived from nonconformity scores calculated over the subsets, each corresponding to different length intervals:

$$\hat{q}_M = \{\hat{q}^i \mid \hat{q}^i = \text{quantile}\left(s_1^i, s_2^i, \ldots, s_{n_i}^i; \lceil (n_i+1)(1-\alpha) \rceil {n_i}^{-1}\right), i = 1, 2, \ldots, K\}, \tag{7}$$

where n_i denotes the number of texts in D_{cal}^i, s^i represents the nonconformity scores calculated from D_{cal}^i, and α denotes the desired upper bound of the FPR. By utilizing $\hat{q}_M$, we can select more appropriate quantiles for calibration across varying text lengths.

3.4 MGT Detection

For a new instance X_{test} from the test set D_{test}, we classify it based on its nonconformity score s_t and text length l_t. The detection result within MCP can be expressed as follows:

$$s_t = (1 + e^{-k(Det(X_{\text{test}}) - \tau)})^{-1}, \tag{8}$$

$$\mathcal{C}(X_{test}) = \mathbb{I}(s_t > \hat{q}^{\lfloor \frac{l_t}{w} \rfloor}), \tag{9}$$

where $\hat{q}^{\lfloor \frac{l_t}{w} \rfloor}$ represents the quantile within the corresponding length interval, and $\mathcal{C}(X_{test})$ denotes the detection result. $\mathcal{C}(X_{test}) = 0$ means that X_{test} is human-written text, while $\mathcal{C}(X_{test}) = 1$ indicates that X_{test} is machine-generated text. A detailed case study is in Appendix A.

Corollary 1. *The upper bound of the FPR for MGT detection within the MCP framework is α. The detailed proof is provided in Appendix B.*

Algorithm 1 MCP Framework

Input: calibration set D_{cal}, test set D_{test}, basic detector Det, nonconformity score function $s(\cdot)$
Compute multiscaled quantiles $\hat{q}_M$:
Number of subsets $K \leftarrow \lfloor \frac{L_{\max}}{w} \rfloor$
Bin calibration set $\{D_{cal}^1, \ldots, D_{cal}^K\} \leftarrow D_{cal}$
for $i = 1$ to K **do**
    $S^i \leftarrow \{s_1^i, \ldots, s_{n_i}^i\}$
    $\delta^i \leftarrow \lceil (n_i+1)(1-\alpha) \rceil {n_i}^{-1}$
    $\hat{q}^i \leftarrow \text{quantile}(S^i; \delta^i)$
end for
$\hat{q}_M \leftarrow \{\hat{q}^i \mid i = 1, 2, \ldots, K\}$
Detect: $X_{\text{test}} \in D_{test}$
$s_t \leftarrow s(X_{\text{test}})$
$\mathcal{C}(X_{test}) \leftarrow \mathbb{I}(s_t > \hat{q}^{\lfloor \frac{l_t}{w} \rfloor})$
if $\mathcal{C}(X_{test}) = 0$ **then**
    X_{test} is a human-written text.
else
    X_{test} is a machine-generated text.
end if

4 RealDet Dataset

Existing datasets are limited in scope and exhibit domain-specific biases (Wu et al., 2024a,b), rendering them inadequate for representing human-written texts across all domains. We introduce the RealDet dataset and compare it with publicly available datasets in Table 1. RealDet offers three key advantages: **(1) Comprehensive Domain Coverage.** RealDet spans **15** distinct textual domains, far exceeding existing datasets in domain diversity. **(2) Extensive Model Coverage.** RealDet is constructed using **22** popular and powerful LLMs, with the broadest range of base models currently. **(3) Large-scale Text Corpus.** RealDet includes over **836k** raw texts (excluding adversarial texts), with more than **106k** human-written texts, significantly surpassing other datasets in the size of raw texts. Furthermore, RealDet includes bilingual texts in both Chinese and English and adversarial texts involving paraphrasing and editing attacks.

Dataset	Origin Size	Domain Coverage	Model Coverage	Multilingual Coverage	Adversarial Coverage
TuringBench (Uchendu et al., 2021)	200k	✗	✓(10)	✗	✗
HC3 (Guo et al., 2023)	26.9k	✓(5)	✗	✓	✗
CHEAT (Yu et al., 2023)	50k	✗	✗	✗	✓
MGTBench (He et al., 2024)	18.5k	✓(3)	✓(5)	✗	✗
M4 (Wang et al., 2024)	122k	✓(5)	✓(7)	✓	✗
MAGE (Li et al., 2024)	447k	✓(10)	✓(9)	✗	✗
RAID (Dugan et al., 2024)	570k	✓(8)	✓(8)	✗	✓
RealDet (Ours)	**836k**	**✓(15)**	**✓(22)**	✓	✓

Table 1: Comparison of open-source datasets in MGT detection. The "Origin size" refers to the number of **raw texts** without adversarial attacks. The "Model Coverage" column represents the count of **base models**.

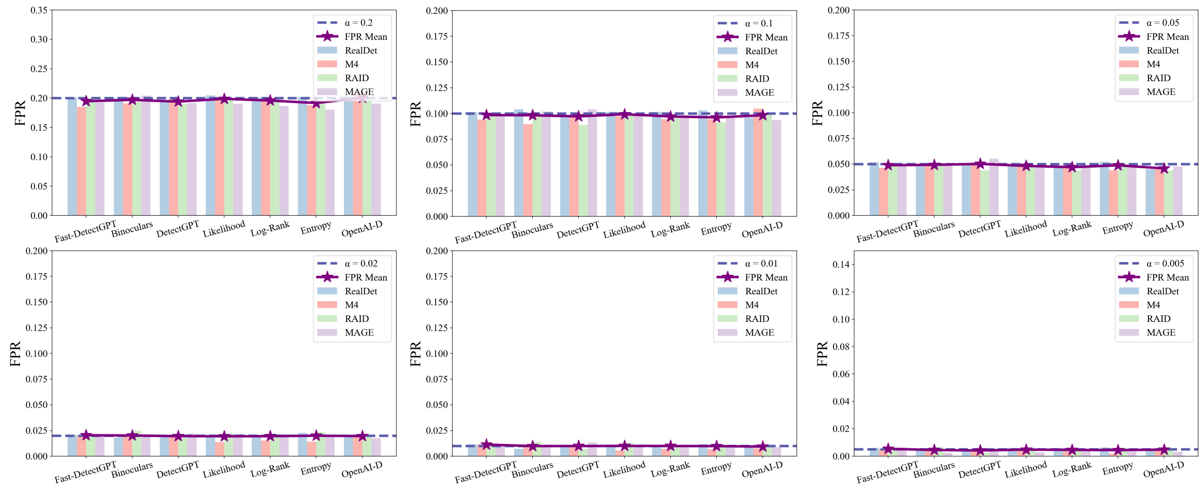

Figure 4: The FPR of various detectors within the MCP framework across all datasets, after applying alpha constraints with values of alpha set to 0.2, 0.1, 0.05, 0.02, 0.01, and 0.005.

Data Sources. To emulate the texts encountered in real-world detection scenarios, we carefully filtered the existing datasets and selected 15 representative data sources covering 6 writing tasks. (1) **Question Answering**: ELI5 (Fan et al., 2019), WiKiQA (Yang et al., 2015), Medical Dialog (He et al., 2020), FiQA (Maia et al., 2018); (2) **News Article Writing**: XSum (Narayan et al., 2018), TLDR[1], BBC News (Greene and Cunningham, 2006); (3) **Story Generation**: WritingPrompt (Fan et al., 2018), ROC Stories (Mostafazadeh et al., 2016); (4) **Review Expression**: Yelp (Zhang et al., 2015), IMDB (Maas et al., 2011), CMV (Tan et al., 2016); (5) **Academic Writing**: Abstracts[2]; (6) **Knowledge Explanation**: Wikipedia (Aaditya Bhat, 2023), SQuAD (Rajpurkar et al., 2016).

Model Set. We consider black-box models (service provider offers API access) and white-box models (open-source models are deployed locally), 22 in total. (1) **Black-box**: Deepseek-V3, GPT-4, GPT-4o, PaLM 2, Ernie Bot 3.5 turbo, Spark Desk 2.0, Qwen turbo, 360GPT S2 V9, Minimax abab 5.5, Claude-3.7 Sonnet; (2) **White-box**: LLaMA2-13B, ChatGLM2-6B, MOSS-moon-003, MPT-7B, InternLM-7B, Alpaca-7B, Guanaco-7B, Vicuna-13B, BLOOMz-7B, Falcon-7B, OPT-6.7B, Baichuan-13B. Details are in Appendix C.1.

Prompt Design. To collect machine-generated text for each instance, we design three types of prompts to feed the LLMs. (1) **Continuation Writing**: ask LLMs to continue generation based on the first sentence of the original human-written text; (2) **Topical Writing**: ask LLMs to generate topic-specific texts (e.g., news article, paper abstract, etc.); (3) **Question-Answering**: ask LLMs to generate an answer based on a given question. Specific prompts are in Appendix C.2.

[1]https://huggingface.co/datasets/JulesBelveze/TLDR_news
[2]https://www.kaggle.com/datasets/spsayakpaul/arxiv-paper-abstracts

5 Experiments

We conduct comprehensive experiments to thoroughly evaluate MCP, focusing on its ability to constrain the FPR, evaluate detection performance, test robustness against real-world attacks, investigate the impact of calibration data, and compare with other calibration methods.

5.1 Experimental Setup

Datasets. We evaluate MCP on RealDet and three representative datasets—M4 (Wang et al., 2024), RAID (Dugan et al., 2024), and MAGE (Li et al., 2024)—all are diverse, high-quality, and large-scale datasets. For each dataset, we randomly sampled 5,000 human-written texts as the calibration set, and 2,500 human-written texts alongside 2,500 machine-generated texts as the test set.

Metrics. We employ the FPR, defined as the proportion of human-written texts misclassified as machine-generated, as the primary metric in MGT detection. Additionally, we use the TPR ($TP@FPR$) and the F1 score ($F_1@FPR$) as metrics to evaluate detection performance.

Basic Detectors. We selected SOTA zero-shot detectors Fast-DetectGPT (Bao et al., 2024) and

Detector	***Algorithm***	$TP@20\%$	$F_1@20\%$	$TP@10\%$	$F_1@10\%$	$TP@5\%$	$F_1@5\%$	$TP@2\%$	$F_1@2\%$	$TP@1\%$	$F_1@1\%$	$TP@0.5\%$	$F_1@0.5\%$
						M4							
Fast-DetectGPT	vanilla	78.56	79.13	74.44	80.72	70.04	80.03	65.32	78.08	60.56	74.97	54.84	70.60
	MCP	**79.56**	**80.35**	**75.24**	**81.50**	**71.72**	**81.33**	**67.36**	**79.55**	**62.44**	**76.33**	**58.64**	**73.67**
Binoculars	vanilla	83.20	81.87	79.44	83.87	74.80	83.19	69.72	81.21	64.52	77.96	55.36	71.03
	MCP	**83.24**	**82.40**	**80.00**	**84.41**	**75.72**	**83.87**	**71.20**	**82.27**	**66.68**	**79.59**	**62.72**	**76.88**
						RAID							
Fast-DetectGPT	vanilla	77.97	78.74	75.23	81.22	71.67	81.13	66.70	79.08	64.70	78.09	63.10	77.14
	MCP	**78.13**	**78.88**	**76.03**	**81.77**	**73.43**	**82.40**	**68.27**	**80.20**	**64.73**	**78.09**	**63.60**	**77.56**
Binoculars	vanilla	78.40	79.03	76.47	82.02	74.53	83.03	70.17	81.50	67.50	80.12	64.17	77.94
	MCP	**78.50**	**79.19**	**76.57**	**82.21**	**74.83**	**83.19**	**72.37**	**82.78**	**70.33**	**81.93**	**66.07**	**79.25**
						MAGE							
Fast-DetectGPT	vanilla	80.36	80.15	77.08	82.40	72.92	81.99	66.04	78.60	57.00	72.15	43.08	60.00
	MCP	**82.12**	**81.57**	**79.72**	**84.08**	**77.28**	**84.83**	**72.44**	**83.02**	**67.92**	**80.40**	**61.24**	**75.64**
Binoculars	vanilla	85.12	83.01	84.00	86.62	82.56	88.04	74.60	84.50	56.04	71.37	28.52	44.20
	MCP	**85.12**	**83.63**	**84.04**	**86.74**	**82.68**	**88.07**	**77.36**	**86.32**	**75.80**	**85.77**	**73.32**	**84.49**
						RealDet							
Likelihood	vanilla	83.60	82.11	79.70	84.03	76.08	84.03	67.38	79.56	58.98	73.73	36.92	53.73
	MCP	**84.10**	**82.23**	**80.64**	**84.53**	**76.66**	**84.50**	**70.06**	**81.43**	**62.24**	**76.21**	**52.60**	**68.67**
Log-Rank	vanilla	84.74	82.80	81.30	84.99	77.96	85.22	70.48	81.73	61.50	75.69	45.74	62.55
	MCP	**85.12**	**82.95**	**81.88**	**85.27**	**78.36**	**85.63**	**72.22**	**82.93**	**65.84**	**78.90**	**58.28**	**73.36**
Entropy	vanilla	68.28	72.51	46.44	59.37	30.24	44.72	14.26	24.53	6.58	12.23	3.18	6.13
	MCP	**70.60**	**74.01**	**48.16**	**60.79**	**31.92**	**46.55**	**17.68**	**29.48**	**9.38**	**16.97**	**5.66**	**10.65**
DetectGPT	vanilla	71.10	74.42	55.14	66.78	38.20	53.35	19.88	32.62	10.42	18.70	5.96	11.19
	MCP	**73.18**	**75.92**	**57.82**	**69.18**	**39.88**	**55.17**	**21.94**	**35.44**	**13.98**	**24.33**	**6.78**	**12.64**
OpenAI-D	vanilla	71.06	74.39	62.94	72.80	57.98	71.16	51.50	67.11	47.32	63.81	43.62	60.53
	MCP	**79.40**	**79.54**	**67.74**	**76.40**	**59.14**	**72.28**	**53.98**	**69.18**	**49.70**	**65.94**	**45.50**	**62.34**
Fast-DetectGPT	vanilla	86.02	83.52	81.86	85.34	77.46	84.91	72.00	82.76	63.74	77.38	51.22	67.52
	MCP	**87.10**	**84.07**	**84.24**	**86.68**	**80.86**	**86.93**	**76.86**	**85.90**	**73.20**	**83.97**	**69.32**	**81.59**
Binoculars	vanilla	90.96	86.20	90.36	90.19	89.16	91.83	84.98	90.90	78.98	87.77	70.16	82.22
	MCP	**91.06**	**86.30**	**90.36**	**90.36**	**89.26**	**92.13**	**87.50**	**92.44**	**86.28**	**92.28**	**84.34**	**91.29**

Table 2: Main Experimental Results Across Various Detectors and Datasets. "Vanilla" refers to the detector's original configuration, with the detection threshold set based on the test set to satisfy a given FPR, whereas "MCP" denotes detectors with MCP framework.

Binoculars (Hans et al., 2024), as well as other zero-shot detectors including DetectGPT (Mitchell et al., 2023), Likelihood, Log-Rank, and Entropy (Gehrmann et al., 2019; Su et al., 2023; Ippolito et al., 2020). We also considered the supervised detectors OpenAI-D (Solaiman et al., 2019b), which utilizes RoBERTa fine-tuned on the GPT-2 dataset. In our experiments, Fast-DetectGPT uses GPT-2-XL as the scoring model and GPT-J-6B as the sampling model; DetectGPT employs GPT-2-XL as the scoring model and T5-3B as the sampling model; Other methods (e.g., Likelihood) uniformly use GPT-2-XL as the scoring model. Considering that different LLMs used as scoring models may result in performance variations, more results are shown in Appendix E.

Hyperparameter Settings. Detailed hyperparameter settings and analysis refer to Appendix F.

5.2 False Positive Rate Constraint

Figure 4 shows the FPR performance of various detectors under the MCP framework across different datasets, with α values selected from $\{0.2, 0.1, 0.05, 0.02, 0.01, 0.005\}$. Each subfigure corresponds to a specific α value, and datasets are distinguished by color. The results confirm that the FPRs are consistently constrained within the theoretical upper bound determined by α, demonstrating MCP's efficacy in controlling false positives. Notably, the uniformity of FPR across detectors highlights the framework's generalizability, making it an effective solution for environments demanding tight FPR control.

5.3 Main Results

Table 2 presents the detection results across four datasets and seven detectors. The MCP framework consistently improves detection performance compared to vanilla detectors, demonstrating strong generalizability. While the gains are modest at higher FPR levels (20%, 10%, and 5%), MCP shows significant improvements at lower FPR thresholds (2%, 1%, and 0.5%). On the RealDet dataset, MCP results in an average improvement of **11%** in TPR and **8%** in F1 score compared to vanilla detectors, and an average improvement of **10%** in TPR and **6%** in F1 score across all datasets. MCP enhances performance by balancing detection accuracy and FPR control through multiscaled conformal quantiles, making it well-suited for applications with strict FPR constraints.

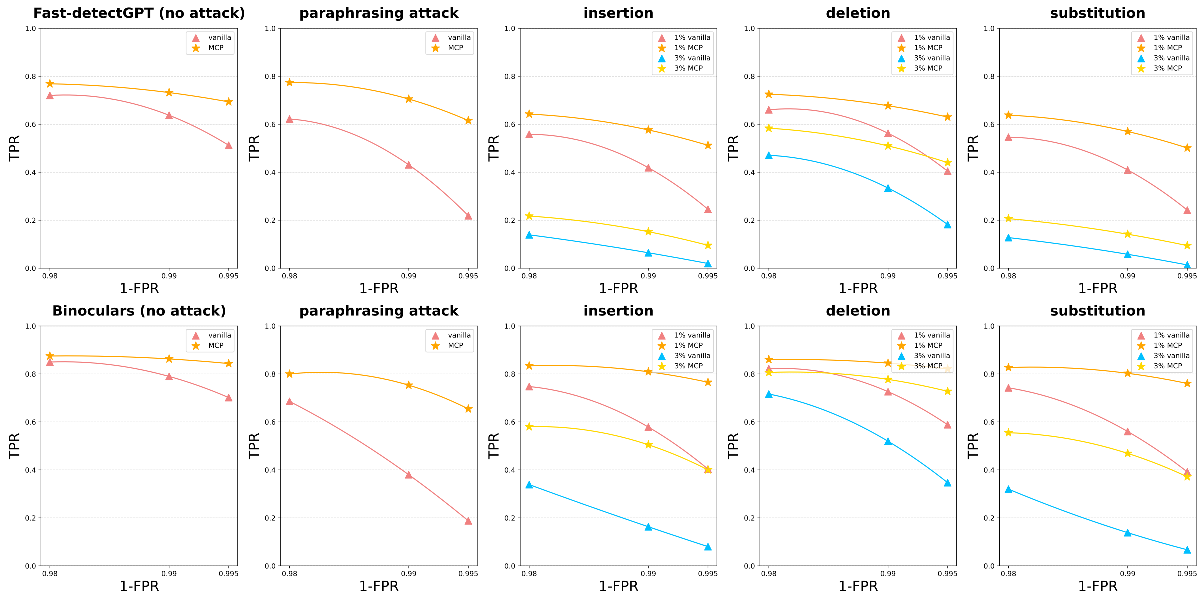

Figure 5: Local ROC curves (with the horizontal axis representing 1 - FPR) for the basic detectors (Binoculars, Fast-DetectGPT) under different real-world attacks, both with and without the MCP framework.

Dataset	*Detector*	*Setting*	TP@1%	F_1@1%	TP@0.5%	F_1@0.5%
MAGE	Fast-DetectGPT	**MCP**	**65.92**	**78.91**	**51.40**	**67.61**
		w/o $\hat{q}_M$	59.76	74.25	48.56	65.08
	Binoculars	**MCP**	**75.80**	**85.77**	**73.32**	**84.49**
		w/o $\hat{q}_M$	50.20	66.49	24.12	38.72
RealDet	Fast-DetectGPT	**MCP**	**73.20**	**83.97**	**69.32**	**81.59**
		w/o $\hat{q}_M$	64.46	77.91	51.72	67.95
	Binoculars	**MCP**	**86.28**	**92.28**	**84.34**	**91.29**
		w/o $\hat{q}_M$	80.82	88.88	73.56	84.44

Table 3: Ablation Study of multiscaled quantiles.

MCP demonstrates superior performance in low-FPR scenarios. For example, on the MAGE dataset, MCP achieves relative improvements of **157%** in TP@0.5% and **91%** in $F1$@0.5%. At higher FPR levels, the improvements are more limited, likely due to the proximity of multiscaled quantiles. Notably, SOTA detectors with MCP maintain high performance under stringent low-FPR constraints. On RealDet, Fast-DetectGPT reaches **69.32%** in TP@0.5% and **81.59%** in $F1$@0.5%, while Binoculars achieves **84.34%** and **91.29%**. MCP's flexibility in adjusting detection thresholds through multiscaled quantiles allows for precise control of low FPRs without sacrificing performance, making it especially effective for high-precision detection.

5.4 Ablation Study

It is important to note that MCP is a framework, and the ablation study focuses on the individual modules within the framework, rather than removing the entire framework itself. Table 3 compares detection performance with and without the multiscaled quantiles calculation module across different datasets. "w/o $\hat{q}_M$" refers to single quantile calculation based on the overall distribution. The results demonstrate that incorporating multiscaled quantiles into the MCP significantly improves detection performance. Specifically, when $\hat{q}_M$ is removed, the average TPR decreases by **22%**, and the average F1 score drops by **15%**. These findings underscore that binning the calibration set and calculating more appropriate quantiles over different length intervals enables more precise calibration, highlighting the necessity of multiscaled quantiles.

5.5 Robustness to Real-world Attacks

Figure 5 illustrates the robustness of the MCP framework under two types of adversarial attacks: paraphrasing and token-level edits (insertion, deletion, and substitution). Paraphrasing attacks were conducted using DIPPER (Krishna et al., 2023) to rephrase the machine-generated texts. Editing attacks involved random insertion, deletion, or substitution of tokens at rates of 1% or 3%.

The results show that, under all attack scenarios, the MCP framework consistently achieves higher TPR compared to the vanilla detectors. Specifically, MCP demonstrates superior resilience, with higher true positive rates across various attack types and intensities. As attack strength increases (from 1% to 3%), MCP continues to maintain a more robust detection performance, whereas vanilla detectors experience more significant drops in TPR. For ex-

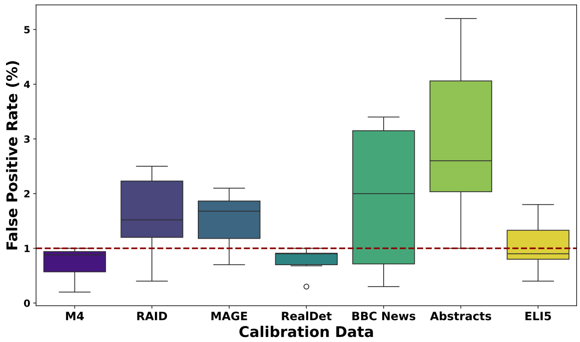

Figure 6: FPR of Binoculars within the MCP at $\alpha = 1\%$ when calibrated with different datasets.

ample, under insertion and deletion attacks, MCP outperforms the vanilla detectors by a substantial margin, indicating its effectiveness in mitigating the impact of adversarial edits. These trends highlight the effectiveness of the MCP framework in maintaining strong detection performance even in challenging adversarial scenarios, emphasizing its robustness compared to the baseline detectors.

5.6 The Impact of Calibration Data

To explore the impact of calibration data on MCP performance, we evaluate multiple datasets, including the multi-domain datasets (M4, RAID, MAGE, RealDet), as well as domain-specific datasets such as news writing (BBC News), academic writing (Abstracts), and social media text (ELI5). As shown in Figure 6, we used a single dataset for calibration, while the remaining datasets were sequentially used as test data.

The use of diverse calibration data generally leads to better performance and improved generalization. Figure 6 shows the FPR of MCP under different calibration data. The results demonstrate that domain-specific datasets (BBC News, Abstracts, ELI5) lead to relatively higher FPRs, suggesting that these datasets are less effective in achieving precise calibration. In contrast, multi-domain datasets (RAID, MAGE) generally perform better, although they still exhibit some limitations due to inherent biases in the data. Notably, M4 and RealDet calibration both yield promising results, with RealDet providing slightly more stable and consistent improvements. Specifically, TP@1% increased by an average of 13% across all test sets when calibrated with RealDet. However, M4 also produces competitive calibration results, highlighting that while RealDet offers a slight edge, diverse calibration data from different domains still plays a critical role in enhancing performance without being overly reliant on a single dataset.

5.7 Comparison with Other Calibration Methods

We compared MCP with other calibration methods in Appendix G, including metric-based maximizing F1 (Lipton et al., 2014), probability distribution-based Platt Scaling (Platt, 1999), and Isotonic Regression(Brunk et al., 1973). While these methods offer a modest improvement in detection, they fall short of effectively controlling the FPR. In contrast, MCP achieves SOTA classification performance while maintaining an exceptionally low FPR, ensuring higher reliability.

6 Related Work

MGT Detection. Existing detectors can be broadly categorized into two main types: zero-shot detectors and supervised detectors. (1) Zero-shot detectors leverage statistical measures extracted by LLMs to identify outliers (Gehrmann et al., 2019; Su et al., 2023; Ippolito et al., 2020; Yang et al., 2023). For instance, the impressive DetectGPT (Mitchell et al., 2023), based on the assumption that MGT is more likely to lie at a local optimum of the log probability, compares log probabilities across multiple perturbations to detect MGT. Fast-DetectGPT (Bao et al., 2024) further improves the text perturbation process of its predecessor, significantly enhancing detection efficiency. Binoculars (Hans et al., 2024) uses cross perplexity between two models from different perspectives to address poor performance when detecting high-perplexity text. (2) supervised detectors typically train a classification model using human-written and machine-generated texts (Solaiman et al., 2019a; Uchendu et al., 2020; Fagni et al., 2021; Zhang et al., 2024; Tian et al., 2024; Pu et al., 2022; Hu et al., 2023; Kumari et al., 2024). Specifically, OpenAI-D (Solaiman et al., 2019b) fine-tuned a RoBERTa model on GPT-2 generated text to detect MGT.

MGT Detection Dataset. Turing Bench (Uchendu et al., 2021) collected 200k human-written texts and machine-generated texts from 19 different models. However, it has become outdated due to the less advanced models. Subsequently, researchers constructed datasets focusing on specific advanced models or particular domains (Fagni et al., 2021; Yu et al., 2023; Mosca et al., 2023). For instance, Guo et al. (2023) built the HC3 dataset by collecting nearly 40k questions covering multiple domains along with corresponding answers generated by human experts and ChatGPT. More recent efforts have introduced large-scale, cross-domain, and cross-model benchmarks, such as the MGTBench (He et al., 2024), M4 (Wang et al., 2024), MAGE (Li et al., 2024), RAID (Dugan et al., 2024), and DetectRL (Wu et al., 2024b) datasets.

7 Conclusion

In this paper, we introduce a reliable machine-generated text detection framework via multiscaled conformal prediction (MCP), which constrains FPRs to mitigate potential societal harms while simultaneously enhancing detection performance. Extensive experiments across seven detectors and four datasets validate the effectiveness of MCP and demonstrate its ability to improve robustness. In future work, we will continue to update our high-quality dataset, RealDet, to address the challenges posed by the rapidly evolving LLMs. Additionally, we plan to investigate an advanced detector, aiming to maintain exceptional detection performance even under stringent FPR within MCP.

Limitations

In our experiments, we found that although we applied a multiscaled optimization strategy using fixed-width binning based on CP, different bin widths consistently corresponded to varying detection performance. Therefore, a more flexible binning strategy could potentially lead to better detection results, an area we have not explored in depth.

Ethics Statement

Detection inherently carries an accusatory implication. While our work constrains the upper bound of the false positive rate, offering more reliable insights to users, we strongly oppose using the detection results from this framework as direct evidence in any punitive context. Regardless of the accuracy of the detection, such use could cause significant harm. Additionally, per the Code of Ethics, no private data or non-public information was used in constructing our dataset.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U2336202).

References

Aaditya Bhat. 2023. Gpt-wiki-intro (revision 0e458f5).

David Ifeoluwa Adelani, Hao Thi Mai, Fuming Fang, Huy Hoang Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection. In *International Conference on Advanced Information Networking and Applications*.

Alim Al Ayub Ahmed, Ayman Aljabouh, Praveen Kumar Donepudi, and Myung Suh Choi. 2021. Detecting fake news using machine learning : A systematic literature review. *Preprint*, arXiv:2102.04458.

M. M. Alshater. 2022. Exploring the role of artificial intelligence in enhancing academic performance: A case study of chatgpt. *SSRN Electronic Journal*.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.

Hugh D. Brunk, Richard E. Barlow, David J. Bartholomew, and Joan M. Bremner. 1973. Statistical inference under order restrictions : the theory and application of isotonic regression. *International Statistical Review*, 41:395.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion

Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jon Christian. 2023. Cnet secretly used AI on articles that didn't disclose that fact, staff say. Accessed: 2023-1-20.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*.

Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Huazuo Gao, Jiashi Li, Liyue Zhang, Panpan Huang, Shangyan Zhou, Shirong Ma, Wenfeng Liang, Ying He, Yuqing Wang, Yuxuan Liu, and Y. X. Wei. 2025. Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures. *Preprint*, arXiv:2505.09343.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Nan Duan. 2016. Overview of the NLPCC-ICCPOL 2016 shared task: Open domain chinese question answering. In *Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2-6, 2016, Proceedings*, volume 10102 of *Lecture Notes in Computer Science*, pages 942–948. Springer.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *PLOS ONE*, 16(5):e0251415.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. *Proceedings of the 23rd international conference on Machine learning*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *Preprint*, arXiv:2301.07597.

Zhiwei Guo, Yu Shen, Ali Kashif Bashir, Muhammad Imran, Neeraj Kumar, Di Zhang, and Keping Yu. 2021. Robust spammer detection using collaborative neural network in internet-of-things applications. *IEEE Internet of Things Journal*, 8(12):9549–9558.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. Mgtbench: Benchmarking machine-generated text detection. *Preprint*, arXiv:2303.14822.

Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. Meddialog: Two large-scale medical dialogue datasets. *Preprint*, arXiv:2004.03329.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *CoRR*, abs/2307.03838.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *CoRR*, abs/2303.13408.

Kavita Kumari, Alessandro Pegoraro, Hossein Fereidooni, and Ahmad-Reza Sadeghi. 2024. DEMASQ: unmasking the chatgpt wordsmith. In *31st Annual Network and Distributed System Security Symposium, NDSS 2024, San Diego, California, USA, February 26 - March 1, 2024*. The Internet Society.

Jing Lei and Larry A. Wasserman. 2014. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Zachary Chase Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. 2014. Thresholding classifiers to maximize f1 score. *Preprint*, arXiv:1402.1892.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla, Paolo Basso, Margherita Musumeci, and Georg Groh. 2023. Distinguishing fact from fiction: A benchmark dataset for identifying machine-generated scientific papers in the LLM era. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 190–207, Toronto, Canada. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alexander Gammerman. 2002. Inductive confidence machines for regression. In *European Conference on Machine Learning*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.

Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abd Ur Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. 2022. Deepfake text detection: Limitations and opportunities. *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1613–1630.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019a. Release strategies and the social impacts of language models. *Preprint*, arXiv:1908.09203.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019b. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *Preprint*, arXiv:2306.05540.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, QINGHUA ZHANG, Ruifeng Li, Chao Xu, and Yunhe Wang. 2024. Multiscale positive-unlabeled detection of AI-generated texts. In *The Twelfth International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vladimir Vovk, Alexander Gammerman, and Craig Saunders. 1999. Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning*.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. Algorithmic learning in a random world.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2024a. A survey on llm-generated text detection: Necessity, methods, and future directions. *Preprint*, arXiv:2310.14724.

Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S. Chao. 2024b. DetectRL: Benchmarking LLM-generated text detection in real-world scenarios. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *Preprint*, arXiv:2305.17359.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. 2023. CHEAT: A large-scale dataset for detecting chatgpt-written abstracts. *CoRR*, abs/2304.12008.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. *Proceedings of the 27th International Conference on Intelligent User Interfaces*.

Shuhai Zhang, Yiliao Song, Jiahao Yang, Yuanqing Li, Bo Han, and Mingkui Tan. 2024. Detecting machine-generated texts by multi-population aware optimization for maximum mean discrepancy. In *The Twelfth International Conference on Learning Representations*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A Case Study

Figure 7 illustrates the detailed detection steps of a single instance within the MCP framework when using Binoculars as the detector, with $w = 100$ and $\alpha = 0.05$. First, the multiscaled quantiles $\hat{q}_M$ are calculated based on the nonconformity score distribution of the calibration data. Next, for a given instance X_{test}, its token length is determined and processed through Binoculars to obtain the output $Det(X_{\text{test}})$. The output $Det(X_{\text{test}})$ is then converted into a nonconformity score s_t following Equation 8. Subsequently, using l_t and w, the corresponding quantile $\hat{q}^0$ from $\hat{q}_M$ is retrieved for the matching length interval. Finally, the prediction result $\mathcal{C}(X_{\text{test}})$ is determined based on the comparison between s_t and $\hat{q}^0$.

B Corollary Proof

This section provides a detailed proof of Corollary 1.

Proof. **Assumptions:**

1. The calibration set D_{cal} and the test set D_{test} are independent and identically distributed (i.i.d.).
2. The conformal prediction framework is employed to calibrate a decision rule based on the calibration set D_{cal}, which is then applied to the test set D_{test}.
3. The nonconformity scores $s(x)$ are properly defined such that higher scores indicate a lower likelihood of the text being HWT.

Within the MCP framework, all of the assumptions above are satisfied.

Objective: To demonstrate that under the MCP framework, the False Positive Rate (FPR) on the test set D_{test} does not exceed the predefined threshold α.

Proof Steps:

Nonconformity Scores Assignment: Assign a nonconformity score $s(x)$ to each instance x in both D_{cal} and D_{test}. These scores quantify how atypical an instance is with respect to the detector.

Multiscaled Quantiles Calculation:

1. Bin the calibration set and sort the nonconformity scores of the subset D^i_{cal} in ascending order:
$$s^i_{(1)} \leq s^i_{(2)} \leq \cdots \leq s^i_{(n)}$$
2. By applying Equation 7 to compute $\hat{q}_M$, the following condition is guaranteed:
$$P(s^i(x) > q^i) \leq \alpha \quad \text{for} \quad x \in D^i_{cal}$$

Bounding the False Positive Rate:

1. Since D_{cal} and D_{test} are i.i.d., the distribution of nonconformity scores in D_{cal} mirrors that of D_{test}. After binning based on text length, and since the selection is performed solely on individual attributes, the subsets (D^i_{cal} and D^i_{test}) obtained for different length intervals remain i.i.d.

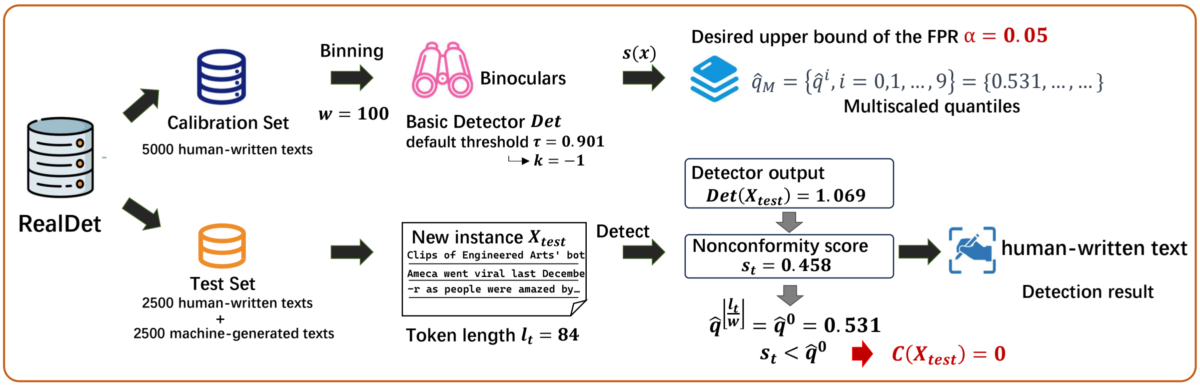

Figure 7: Detailed Detection Process of a Single Instance within the MCP Framework.

2. Consequently, the same $\hat{q}_M$, when applied to the test set, continues to satisfy the following condition:

$$P(s^i(x) > q^i) \leq \alpha \quad \text{for} \quad x \in D^i_{test}$$

3. The FPR is computed from the instances across all length intervals in D_{test}:

$$\text{FPR} = \mathbb{E}\left(P\left(s^i(x) > q^i \mid i = 1, 2, \ldots\right)\right) \leq \alpha$$

□

C Details of RealDet

In this subsection, we will provide additional detailed information regarding various aspects of RealDet.

C.1 Model Supplement

Table 4 presents all LLMs along with their corresponding text quantities. In collecting machine-generated texts, we considered two distinct settings: black-box models and white-box models. From the black-box models, we selected 10 popular and powerful LLMs for data collection, namely Deepseek-V3 (Zhao et al., 2025), GPT-4, GPT-4o, PaLM 2, Ernie Bot 3.5 Turbo, Spark Desk 2.0, Qwen Turbo, 360GPT S2 V9, Minimax Abab 5.5, and Claude-3.7 Sonnet. Similarly, from the white-box models, we selected 12 LLMs, including LLaMA2-13B (Touvron et al., 2023), ChatGLM2-6B (GLM et al., 2024), MOSS-moon-003, MPT-7B (Team, 2023), InternLM-7B (Cai et al., 2024), Alpaca-7B (Taori et al., 2023), Guanaco-7B (Dettmers et al., 2023), Vicuna-13B (Chiang et al., 2023), BLOOMz-7B (Muennighoff et al., 2023), Falcon-7B (Penedo et al., 2023), OPT-6.7B (Zhang et al., 2022), and Baichuan-13B. The number of machine-generated English texts exceeds 604k, whereas the number of machine-generated Chinese texts surpasses 125k. The total number of texts exceeds 836k.

Source Model	En-Text	Cn-Text	Total
Deepseek-V3	20,783	12,784	33,567
GPT-4	14,307	-	14,307
GPT-4o	20,600	12,376	32,976
PaLM2	21,878	-	21,878
Ernie Bot turbo 3.5	45,959	12,011	57,970
Spark Desk 2.0	42,963	8,280	51,243
Qwen turbo	45,911	11,993	57,904
360GPT S2 V9	44,833	12,054	56,887
Minimax abab 5.5	19,954	-	19,954
Claude-3.7 Sonnet	20,792	12,834	33,626
LLaMA2-13B	33,486	-	33,486
ChatGLM2-6B	29,132	12,221	41,353
MOSS-moon-003	28,902	10,721	39,623
MPT-7B	29,287	-	29,287
InternLM-7B	16,435	-	16,435
Alpaca-7B	28,233	-	28,233
Guanaco-7B	29,034	-	29,034
Vicuna-13B	28,520	-	28,520
BLOOMz-7B	7,259	7,782	15,041
Falcon-7B	22,501	-	22,501
OPT-6.7B	25,159	-	25,159
Baichuan-13B	29,060	12,239	41,299
Human	96,150	10,207	106,357
Total	**701,138**	**135,502**	**836,640**

Table 4: Specific Quantities in different LLMs generated texts.

C.2 Prompt Design

In this study, we designed 3 types of generic prompts—continuation writing, topical writing, and question-answering—for LLMs to generate texts across 15 distinct domains. Table 11 presents sample prompts for the continuation type, Table 12 showcases examples of thematic writing prompts, and Table 13 illustrates examples of question-

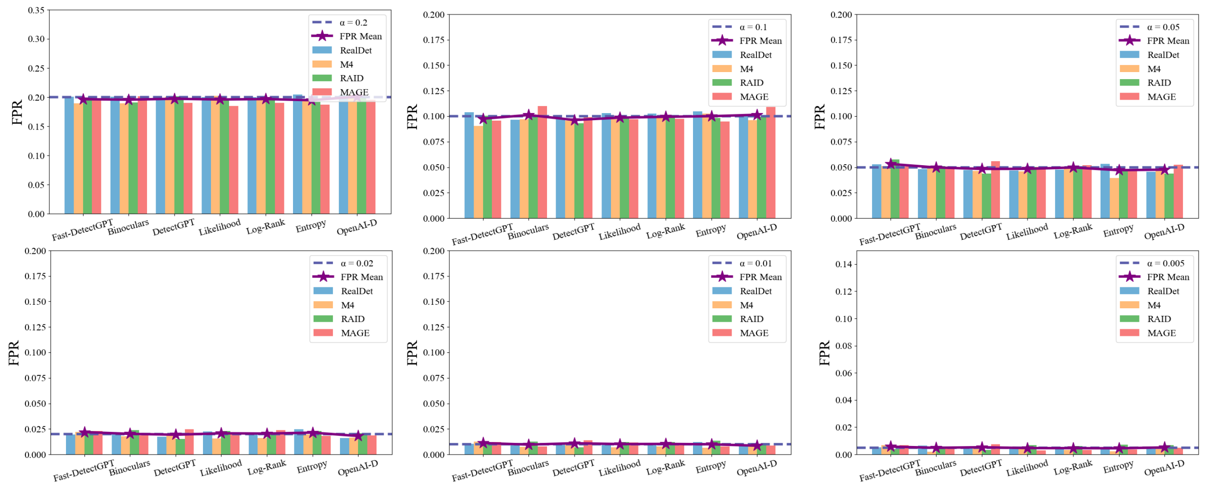

Figure 8: The FPR of various detectors within Traditional CP framework across all datasets, after applying alpha constraints with values of alpha set to 0.2, 0.1, 0.05, 0.02, 0.01, and 0.005.

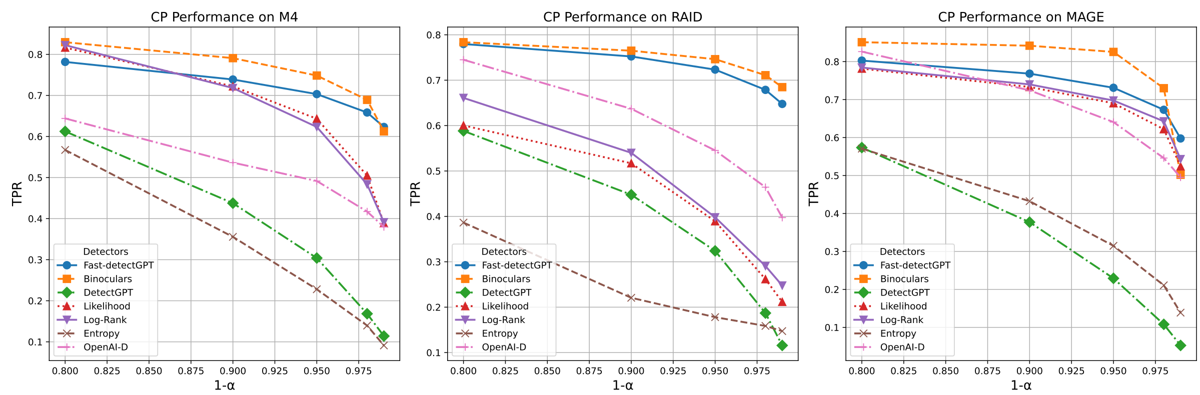

Figure 9: The TPR of various detectors within Traditional CP framework across 3 datasets.

answering prompts.

C.3 Adversarial Attacks

In this study, we do not consider adversarial attacks on human-written texts, as evading detection with human-written content is deemed inconsequential. Instead, we focus on adversarial attacks in the context of machine-generated texts by introducing 2 common attack types: paraphrasing attacks and editing attacks. For paraphrasing attacks, we employ DIPPER with hyperparameters set to a lexical diversity of 60 and a syntactic diversity of 60. This level of paraphrasing is sufficient to potentially bypass state-of-the-art (SOTA) detectors. Regarding editing attacks, we utilize the GPT-2 tokenizer to encode the text and obtain a token sequence. We then apply random insertions, deletions, and substitutions to the token sequence at proportions of 1%, 3%, and 5%. The tokens inserted and substituted are randomly selected from the tokenizer's vocabulary. In the adversarial attack scenario, each raw text is associated with 10 adversarial texts (calculated as 1 original + 3 proportions × 3 types of edits).

D Performance of the Traditional CP

Figure 8 shows the FPR performance of different detectors under the traditional CP framework across various datasets. We observe that the traditional CP framework effectively constrains the upper bound of the FPR. Figure 9 presents the TPR performance of different detectors within the traditional CP framework. Although the traditional CP framework successfully limits the FPR upper bound, we find that the TPR sharply decreases under low FPR settings. Even SOTA detectors experience a significant decline in TPR, making it difficult to detect MGTs. This observation serves as the motivation for our proposed MCP approach.

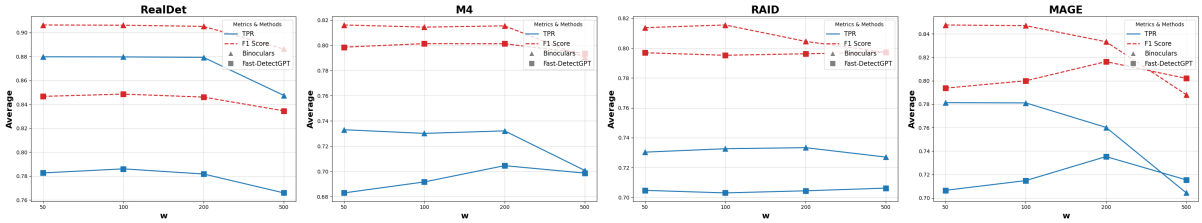

Figure 10: Hyperparameter Sensitivity Analysis of w.

E More Supplement Results

Consistent Positive Correlation. Table 5 presents the quantile values of various detectors across different text lengths and Pearson correlation coefficients $\rho_{(l,s)}$ (obtained by quantile and average interval length) when $\alpha = 0.01$. It is evident that the quantile values of different detectors show a consistent positive correlation with text length.

Performance with the same scoring model. Considering that different LLMs used as scoring models may result in performance variations, Table 6 further includes experimental comparisons with Falcon-7B-instruct uniformly set as the scoring model and Falcon-7B as the sampling model, thus eliminating the impact introduced by different LLMs. The results demonstrate that MCP framework brings consistent performance improvements, aligned with the main results of the paper, particularly under settings with low false positive rates. Specifically, Binoculars within the MCP framework exhibits greater performance gains compared to Fast-DetectGPT and achieves comparable performance when using the same scoring model.

F Hyperparameter Sensitivity Analysis

The MCP framework encompasses 3 hyperparameters: the upper bound on the FPR (α), the maximum input length (L_{max}), and the bin width (w). In the main experiments, the upper bound α is typically selected from the set $\{0.2, 0.1, 0.05, 0.02, 0.01, 0.005\}$. L_{max} is contingent upon the basic detector integrated within the framework. For instance, the OpenAI-D detector, which is fine-tuned based on RoBERTa, utilizes $L_{\text{max}} = 512$, whereas Fast-DetectGPT, which employs GPT-2 for sampling and scoring, adopts $L_{\text{max}} = 1024$. The bin width w is influenced by both the data distribution and the output distribution. In the main experiments, a default bin width of $w = 100$ is employed, as this width yields the most optimal and stable performance within the framework.

Figure 10 presents the performance of varying bin width (w) values across different datasets and detectors. We observe that performance remains consistently strong when $w \leq 200$. However, at $w = 500$, detection performance noticeably declines, indicating that excessively large bin widths undermine the MCP framework's ability to enhance detection performance.

The optimal bin width is influenced by the distribution of the textual data. Specifically, within the M4 dataset, a bin width of $w = 200$ consistently outperforms other values in detection performance. We attribute this to the inherent text distribution of the M4 dataset, where the majority of texts pertain to academic writing and peer review, introducing a certain degree of bias.

The distribution of the detector's outputs affects the optimal selection of w. The original output distribution of Binoculars is more concentrated, leading to a similarly concentrated distribution of the transformed nonconformity scores. Consequently, a smaller and more refined bin width enhances detection performance. In contrast, Fast-DetectGPT exhibits a relatively dispersed and smooth output distribution, suggesting that a larger bin width may achieve superior detection performance in this context.

G Comparison with Other Calibration Methods

Existing calibration methods can be categorized into two main types: metric-based methods (e.g., maximizing F1) and probability distribution-based methods (e.g., Platt Scaling and Isotonic Regression). Maximizing F1 derives the optimal threshold by identifying the threshold that maximizes the F1 score, whereas Platt Scaling and Isotonic Regression modify the output probability distribution to enhance performance. MCP differs from other calibration methods in the following two key aspects:

Detector	*Length*	(0,100)	[100,200)	[200,300)	[300,400)	[400,500)	[500,600)	[600,700)	[700,800)	[800,900)	[900,1024]	$\rho_{(l,s)}$
Likelihood	quantile	0.3516	0.3481	0.3448	0.3545	0.3603	0.3660	0.3786	0.3984	0.4582	0.5440	0.8200
LogRank	quantile	0.3899	0.3854	0.3771	0.3818	0.3802	0.3877	0.3915	0.4052	0.4157	0.4585	0.7533
Entropy	quantile	0.0591	0.1268	0.1492	0.1441	0.1575	0.1647	0.1617	0.1724	0.1656	0.1636	0.7622
Binoculars	quantile	0.4089	0.4410	0.4491	0.4519	0.4577	0.4588	0.4583	0.4588	0.4571	0.4578	0.7344
Fast-DetectGPT	quantile	0.6788	0.6826	0.6781	0.7540	0.8093	0.8577	0.8724	0.9400	0.9721	0.9920	0.9842

Table 5: Quantiles Across Different Lengths for Various Detectors

Detector	***Algorithm***	$TP@2\%$	$F_1@2\%$	$TP@1\%$	$F_1@1\%$	$TP@0.5\%$	$F_1@0.5\%$
				RealDet			
Likelihood	vanilla	72.72	83.24	59.46	74.11	37.76	54.62
	MCP	**74.88**	**84.71**	**62.96**	**76.84**	**48.34**	**65.02**
Log-Rank	vanilla	74.74	84.58	64.50	77.95	49.80	66.27
	MCP	**75.64**	**85.26**	**65.66**	**78.88**	**56.94**	**72.32**
Entropy	vanilla	65.00	77.84	49.84	66.08	31.52	47.75
	MCP	**67.52**	**79.71**	**54.34**	**70.06**	**40.10**	**57.06**
Fast-DetectGPT	vanilla	88.42	92.87	86.88	92.47	84.08	91.10
	MCP	**88.84**	**92.98**	**87.60**	**92.88**	**86.26**	**92.37**
Binoculars	vanilla	84.98	90.90	78.98	87.77	70.16	82.22
	MCP	**87.50**	**92.44**	**86.28**	**92.28**	**84.34**	**91.29**

Table 6: Supplementary experimental results on RealDet, with Falcon-7B-instruct as the scoring model and Falcon-7B as the sampling model.

Detector	***Method***	FPR	F_1
Fast-DetectGPT	**vanilla**	9.84	85.36
	Maximizing F_1	9.24	85.53
	Platt Scaling	8.46	85.65
	Isotonic Regression	8.78	85.58
	MCP ($\alpha = 10\%$)	**9.94**	**86.68**
	MCP ($\alpha = 5\%$)	**4.98**	**86.93**
	MCP ($\alpha = 2\%$)	**1.98**	**85.90**
Binoculars	**vanilla**	7.68	91.06
	Maximizing F_1	7.26	91.21
	Platt Scaling	6.54	91.43
	Isotonic Regression	7.34	91.18
	MCP ($\alpha = 5\%$)	**4.78**	**92.13**
	MCP ($\alpha = 2\%$)	**1.82**	**92.44**
	MCP ($\alpha = 1\%$)	**0.72**	**92.28**

Table 7: Detection Performance with Different Calibration Methods.

(1) MCP effectively constrains the upper bound of the FPR, whereas other calibration methods primarily optimize classification performance without explicitly controlling FPR. (2) MCP is a zero-shot, dynamic threshold calibration method. Maximizing F1 determines a fixed threshold that maximizes F1 performance, while Platt Scaling and Isotonic Regression train a calibration model to optimize the output probability distribution and then classify instances based on the newly calibrated probabilities. In contrast, MCP requires no additional training and dynamically adjusts the threshold across different text lengths, demonstrating its efficiency and flexibility.

Table 7 compares the experimental performance of MCP with other calibration methods. Using the RealDet dataset, we conducted experiments on Fast-DetectGPT and Binoculars with different calibration methods. The results indicate that although other methods (maximizing F1, Platt Scaling, and Isotonic Regression) achieve modest improvements in classification performance, they still exhibit dangerously high FPRs (FPR > 5%). In contrast, MCP achieves SOTA classification performance and maintains an exceptionally low FPR (FPR < 2%), ensuring higher reliability.

H Main Experiment Supplement

Tables 8, 9, and 10 present a comparative analysis of the detection performance of 7 different detectors on the M4, RAID, and MAGE datasets, respectively, both with and without the MCP framework. These tables serve as supplementary material to the primary experiments. The results consistently demonstrate that the MCP framework enhances the detection capabilities of the detectors, particularly under settings with low false positive rates. Furthermore, it was observed that non-state-of-the-art (non-SOTA) detectors sometimes do not exhibit performance improvements when integrated with the MCP framework, especially when their baseline performance is bad. We attribute this phenomenon to the possibility that inaccurate outputs from non-SOTA detectors may hinder the calibration process within the MCP framework. This observation aligns with the conclusions drawn in Subsection 5.3, where it was noted that accurate outputs from SOTA detectors facilitate more effective calibration within MCP.

Detector	***Algorithm***	TP@20%	F_1@20%	TP@10%	F_1@10%	TP@5%	F_1@5%	TP@2%	F_1@2%	TP@1%	F_1@1%	TP@0.5%	F_1@0.5%
							M4						
Likelihood	vanilla	81.48	80.87	72.76	79.62	65.12	76.54	54.72	69.85	43.92	60.61	35.96	52.70
	MCP	87.28	84.17	78.68	83.52	70.04	80.10	58.12	72.89	48.28	64.88	36.04	52.83
Log-Rank	vanilla	82.52	81.49	71.84	79.03	62.04	74.28	50.76	66.47	42.96	59.68	35.32	52.02
	MCP	87.24	84.32	78.80	83.72	70.12	80.12	56.68	71.66	48.48	64.99	38.56	55.51
Entropy	vanilla	57.12	64.50	35.28	48.53	25.12	38.61	14.00	24.14	10.48	18.80	8.48	15.57
	MCP	64.80	70.59	45.48	58.71	28.44	42.82	14.84	25.53	9.88	17.87	6.48	12.14
DetectGPT	vanilla	61.20	67.53	44.68	57.77	31.12	45.72	16.96	28.51	10.08	18.15	7.28	13.51
	MCP	62.00	68.19	46.28	59.35	31.52	46.12	18.36	30.53	11.52	20.49	7.08	13.17
OpenAI-D	vanilla	64.00	69.53	54.12	65.95	49.60	64.17	42.60	58.92	38.56	55.26	35.88	52.63
	MCP	67.24	71.43	59.28	69.84	52.08	66.45	45.04	61.28	38.92	55.73	35.04	51.76
Fast-DetectGPT	vanilla	78.56	79.13	74.44	80.72	70.04	80.03	65.32	78.08	60.56	74.97	54.84	70.60
	MCP	**79.56**	**80.35**	**75.24**	**81.50**	**71.72**	**81.33**	**67.36**	**79.55**	**62.44**	**76.33**	**58.64**	**73.67**
Binoculars	vanilla	83.20	81.87	79.44	83.87	74.80	83.19	69.72	81.21	64.52	77.96	55.36	71.03
	MCP	**83.24**	**82.40**	**80.00**	**84.41**	**75.72**	**83.87**	**71.20**	**82.27**	**66.68**	**79.59**	**62.72**	**76.88**

Table 8: Main Experimental Supplement on the **M4** dataset. "Vanilla" refers to the detector's original configuration, whereas "MCP" denotes detectors with MCP framework.

Detector	***Algorithm***	TP@20%	F_1@20%	TP@10%	F_1@10%	TP@5%	F_1@5%	TP@2%	F_1@2%	TP@1%	F_1@1%	TP@0.5%	F_1@0.5%
							RAID						
Likelihood	vanilla	64.37	69.82	52.47	64.59	38.70	53.86	24.80	39.12	19.37	32.18	14.00	24.46
	MCP	62.93	68.87	53.57	65.59	44.90	60.00	36.50	52.63	32.17	48.20	26.63	41.83
Log-Rank	vanilla	66.10	71.01	54.10	65.91	40.63	55.80	28.83	44.08	22.67	36.66	19.77	32.88
	MCP	64.97	70.26	54.73	66.60	46.47	61.61	39.43	55.70	34.23	50.59	29.40	45.23
Entropy	vanilla	40.03	50.06	22.37	33.80	17.90	29.14	15.57	26.48	14.13	24.54	11.77	20.96
	MCP	42.33	52.30	29.33	42.37	23.53	36.61	17.43	29.11	14.73	25.43	12.27	21.74
DetectGPT	vanilla	58.57	65.62	46.53	59.45	34.50	49.46	21.20	34.41	14.67	25.36	8.80	16.10
	MCP	60.10	67.06	48.37	61.51	35.20	50.64	21.20	34.54	12.67	22.38	7.47	13.86
OpenAI-D	vanilla	74.47	76.57	63.80	73.40	56.07	69.62	45.63	61.82	39.93	56.67	17.00	28.94
	MCP	73.83	76.33	62.63	72.87	54.77	68.83	47.07	63.13	41.93	58.66	33.47	49.93
Fast-DetectGPT	vanilla	77.97	78.74	75.23	81.22	71.67	81.13	66.70	79.08	64.70	78.09	63.10	77.14
	MCP	**78.13**	**78.88**	**76.03**	**81.77**	**73.43**	**82.40**	**68.27**	**80.20**	**64.73**	**78.09**	**63.60**	**77.56**
Binoculars	vanilla	78.40	79.03	76.47	82.02	74.53	83.03	70.17	81.50	67.50	80.12	64.17	77.94
	MCP	**78.50**	**79.19**	**76.57**	**82.21**	**74.83**	**83.19**	**72.37**	**82.78**	**70.33**	**81.93**	**66.07**	**79.25**

Table 9: Main Experimental Supplement on the **RAID** dataset. "Vanilla" refers to the detector's original configuration, whereas "MCP" denotes detectors with MCP framework.

Detector	***Algorithm***	TP@20%	F_1@20%	TP@10%	F_1@10%	TP@5%	F_1@5%	TP@2%	F_1@2%	TP@1%	F_1@1%	TP@0.5%	F_1@0.5%
							MAGE						
Likelihood	vanilla	78.52	79.12	73.60	80.14	68.88	79.23	61.56	75.28	49.84	66.08	37.20	54.02
	MCP	79.56	80.14	74.60	80.81	69.20	79.50	62.96	76.26	53.08	68.88	35.44	52.21
Log-Rank	vanilla	78.64	79.18	74.12	80.43	69.48	79.62	61.64	75.34	51.00	67.11	45.52	62.34
	MCP	79.28	80.11	74.52	80.95	70.24	80.24	64.24	77.17	55.00	70.51	44.60	61.53
Entropy	vanilla	58.96	65.88	44.76	57.84	31.76	46.43	22.68	36.38	15.44	26.52	10.52	18.95
	MCP	58.76	66.47	43.20	56.63	31.72	46.46	21.44	34.85	16.12	27.55	10.44	18.84
DetectGPT	vanilla	59.36	66.21	38.20	51.54	19.80	31.73	7.64	13.94	3.88	7.40	2.04	3.98
	MCP	58.16	65.60	39.92	53.11	24.76	38.01	11.40	20.07	6.36	11.81	2.92	5.64
OpenAI-D	vanilla	82.84	81.68	71.72	78.92	63.08	75.06	54.96	70.03	49.96	66.19	43.08	60.02
	MCP	80.96	80.96	74.04	80.74	67.68	78.50	59.36	73.68	53.16	69.02	44.76	61.70
Fast-DetectGPT	vanilla	80.36	80.15	77.08	82.40	72.92	81.99	66.04	78.60	57.00	72.15	43.08	60.00
	MCP	**82.12**	**81.57**	**79.72**	**84.08**	**77.28**	**84.83**	**72.44**	**83.02**	**67.92**	**80.40**	**61.24**	**75.64**
Binoculars	vanilla	85.12	83.01	84.00	86.62	82.56	88.04	74.60	84.50	56.04	71.37	28.52	44.20
	MCP	**85.12**	**83.63**	**84.04**	**86.74**	**82.68**	**88.07**	**77.36**	**86.32**	**75.80**	**85.77**	**73.32**	**84.49**

Table 10: Main Experimental Supplement on the **MAGE** dataset. "Vanilla" refers to the detector's original configuration, whereas "MCP" denotes detectors with MCP framework.

Prompt Type	Prompt	Generated Text	Domain	Model
Continuation Writing	First, I'm well aware of the fiasco that was 1989 on the cold fusion front. Please continue.	Certainly, after Martin Fleischmann and Stanley Pons announced they had achieved cold fusion in 1989, many laboratories across the world attempted to replicate these results with mixed results. This brought the legitimacy of their claims into question and the scientific community largely dismissed cold fusion as a viable technology.	CMV	GPT-4
	While the main story is supposed to take place in Morocco, this movie was shot in foggy Romania in 18 days on a very tight budget. Please continue.	The decision to film in Romania instead of Morocco was likely influenced by financial constraints and logistical challenges. Shooting in a foreign country can be expensive, especially for a production with a limited budget. Romania offered more cost-effective options, including lower labor costs, tax incentives, and accessible filming locations that could double for Moroccan settings.	IMDB	360GPT S2 V9
	In 1986, Michael Dell brought in Lee Walker, a 51-year-old venture capitalist, as president and chief operating officer, to serve as Michael's mentor and implement Michael's ideas for growing the company. Please continue.	With Lee Walker's appointment as President and Chief Operating Officer in 1986, Dell Computer Corporation began to experience significant growth under Michael Dell's vision and Walker's experienced guidance. Walker, a seasoned venture capitalist, brought a wealth of knowledge and expertise in finance, operations, and strategic planning to the company.	SQuAD	Ernie Bot 3.5 turbo

Table 11: Some Examples in Continuation Writing.

Prompt Type	Prompt	Generated Text	Domain	Model
Topical Writing	Please write an abstract based on the following title “Addressing Visual Search in Open and Closed Set Settings”.	Certainly! Here is an abstract based on the title “Addressing Visual Search in Open and Closed Set Settings”: In recent years, visual search has become a critical component of various applications, ranging from e-commerce to augmented reality. This paper explores the challenges and solutions associated with visual search in both open and closed set settings. In closed set scenarios, where all possible categories are known, the focus is on accurate classification and retrieval, ensuring high precision and recall rates. Conversely, open set scenarios involve the identification of unknown categories, necessitating robust mechanisms for distinguishing novel objects from familiar ones. We present a comprehensive framework that leverages deep learning models tailored for each setting, incorporating strategies such as feature extraction, similarity metrics, and anomaly detection. Our approach aims to enhance the performance of visual search systems by adapting to the dynamic nature of real-world data, thereby bridging the gap between theoretical advancements and practical applications. Experimental results demonstrate significant improvements in accuracy and efficiency, paving the way for more versatile and reliable visual search solutions.	Abstracts	Qwen turbo
	Please write a news article based on the following title: “Wales hails new superstar”.	Title: Wales Hails New Superstar Subtitle: The Rise of a Welsh Sporting Icon Date: 2023-04-10 Wales has been abuzz with excitement as a new superstar emerges in the world of sports. The unveiling of this extraordinary talent has captured the hearts and imaginations of the Welsh public and has sparked a renewed sense of pride and optimism for the future of Welsh sports. The young athlete, who wishes to remain anonymous for the time being, has been making waves in their chosen sport, demonstrating an unparalleled level of skill, determination, and sportsmanship. Their meteoric rise to fame has been nothing short of remarkable, with many industry experts predicting a long and illustrious career ahead. The new superstar’s impact has been felt far beyond the realm of sports, with their success inspiring a new generation of Welsh athletes to pursue their dreams. Their story serves as a powerful reminder of the importance of hard work, dedication, and perseverance in achieving one’s goals. As the excitement surrounding this new superstar continues to grow, the Welsh public eagerly awaits their next move. With the eyes of the world now fixed on Wales, there is no doubt that the future of Welsh sports is in good hands. In conclusion, the emergence of this new superstar has injected a much-needed dose of enthusiasm and optimism into the Welsh sports scene. Their remarkable journey serves as an inspiration to us all, and we eagerly await the next chapter in their incredible story.	BBC News	Minimax abab 5.5

Table 12: Some Examples in Topical Writing.

Prompt Type	Prompt	Generated Text	Domain	Model
Question-Answering	Please explain what is "2021 FIBA Women's AmeriCup".	The 2021 FIBA Women's AmeriCup is a basketball tournament for women's national teams from the Americas, organized by the International Basketball Federation (FIBA). It represents one of FIBA Americas' biggest events, featuring top teams from across the continent. In 2021, the event took place from June 11 to June 19 in San Juan, Puerto Rico. Ten teams participated in the tournament, which aimed to provide international competition experience for players and contribute to the global promotion of the sport. The United States won the tournament, defeating Puerto Rico in the final.	Wikipedia	GPT-4
	How can I judge loan availability?	To judge loan availability, you can consider the following factors: Credit Score: Your credit score is a key factor in determining loan availability. Lenders use credit scores to assess the risk of lending money to you. A good credit score indicates that you are a low-risk borrower and may qualify for better loan terms. Income: Lenders will consider your income level when determining loan availability. They will want to ensure that you have the ability to repay the loan. Debt-to-Income Ratio: Your debt-to-income ratio is a measure of how much of your income is going towards debt payments. Lenders will consider this ratio when evaluating loan applications. A high debt-to-income ratio may indicate that you have too much debt and may not be a good candidate for a loan. Credit History: Your credit history is another important factor in determining loan availability. Lenders will consider how long you have had credit, whether you have any negative marks on your credit report, and how you have managed your credit in the past. Loan Type: The type of loan you are applying for will also impact its availability. Different types of loans, such as personal loans, mortgages, or auto loans, have different requirements and conditions. By considering these factors, you can get a better idea of whether a loan is available to you and what terms you may be offered.	FiQA	GPT-4

Table 13: Some Examples in Question-Answering.