WHAT HAPPENS AFTER THE SHIFT? TRAINING-FREE DETECTION OF LLM-GENERATED TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

The increasing use of large language models (LLMs) has created a growing need for reliable methods to distinguish human-written from LLM-generated text. Training-free detectors are attractive because they do not require labeled training data or access to the source generator, but existing approaches primarily rely on global likelihood statistics, temporal fluctuations, or spectral characteristics of token probability sequences. In this work, we investigate whether local transition behavior in token log-probability sequences provides an additional discriminative signal. We identify the dominant token-to-token transition and characterize the variability of the likelihood trajectory following this transition. Based on these observations, we propose ExpStabilityDet, a training-free detector that combines global likelihood statistics with transition-aware features, including the dominant transition magnitude and post-transition variability. We further introduce ExpStabilityDet++, a sampling-based extension that evaluates the observed text relative to conditionally generated contrast texts. Across multiple source models, datasets, response lengths, decoding strategies, and languages, our approach consistently demonstrates the usefulness of transition-aware likelihood characteristics. In the primary black-box evaluation, ExpStabilityDet achieves an average AUROC of 0.8877, while ExpStabilityDet++ achieves 0.9157, compared with 0.8785 and 0.8881 for SpecDetect and SpecDetect++, respectively. Additional experiments show that the proposed representation remains informative under cross-lingual and paraphrasing settings and provides a favorable performance–runtime trade-off. These results suggest that the relationship between a salient likelihood transition and its subsequent trajectory provides complementary information for training-free LLM-generated text detection [1].

1 INTRODUCTION

The rapid advancement of large language models (LLMs) has fundamentally transformed natural language generation. Modern AI systems such as LLaMA Adcock et al. (2026), GPT-4 Singh et al. (2025), and Qwen Yang et al. (2025) are capable of producing highly coherent, fluent, and contextually sophisticated text that increasingly resembles human writing, raising urgent concerns regarding misinformation and academic integrity Park & Nan (2025); Susnjak & McIntosh (2024). To address these risks, substantial research has focused on detecting LLM-generated text. Existing approaches can be broadly categorized into training-based and training-free. While training-based methods Adam et al. (2026); Liu et al. (2023); Guo et al. (2023) can be effective, they often require substantial labeled data and lack to generalize across unseen domains or models. In contrast, training-free methods Gehrmann et al. (2019); Yang et al.; Hans et al. leverage intrinsic statistical properties, such as token likelihoods and entropy, offering stronger cross-model and cross-domain robustness without expensive labeled datasets.

Recent training-free methods have evolved from simple global statistics toward more sophisticated modeling of token probability dynamics. Lastde Xu et al. treats the token probability sequence as a time series to capture local fluctuations through diversity entropy, while SpecDetect Luo et al. (2026) utilizes the frequency domain to measure generative vitality through spectral energy. In a parallel direction, DNA-DetectLLM Zhu et al. (2025) introduces the frequency-domain characteristics of

[1] The code and datasets used in this work will be made publicly available upon acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

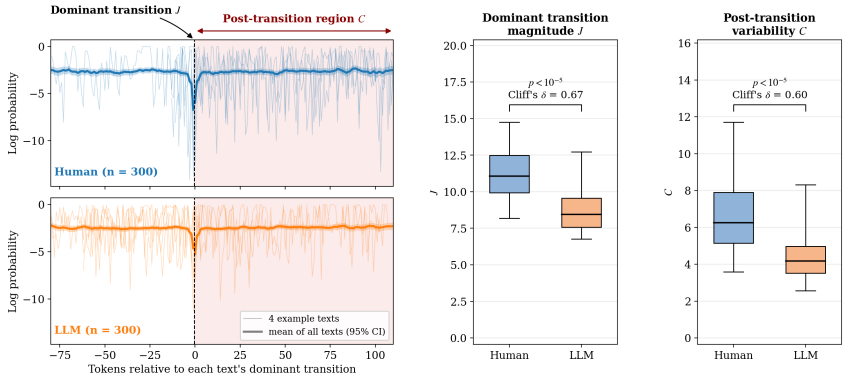


Figure 1: Empirical motivation for transition-aware likelihood analysis. Token log-probability sequences are aligned at the dominant transition of each text. The left panel illustrates the likelihood trajectories around the dominant transition, while the middle and right panels show the distributions of the dominant transition magnitude J and post-transition variability C for human-written and LLM-generated texts, respectively.

token probability sequences. These developments indicate that the sequence of token likelihoods contains information beyond a single aggregate likelihood value. Nevertheless, existing approaches primarily characterize global statistics, aggregate temporal fluctuations, or spectral properties. They do not explicitly examine how a salient local transition in the likelihood trajectory relates to the behavior that follows that transition.

To investigate this overlooked aspect, we perform an empirical analysis of token log-probability sequences from human-written and LLM-generated texts. For each text, we identify its dominant transition, i.e., the largest token-to-token change in log-probability, and align the likelihood trajectory around this transition. Figure 1 illustrates the resulting patterns. The likelihood trajectories exhibit visibly different behavior around the dominant transition. Moreover, the distributions of the dominant transition magnitude J and the variability of the post-transition region C differ substantially between human-written and LLM-generated texts. In our analysis, both differences are statistically significant ($p < 10^{-5}$), with Cliff's δ values of 0.67 for J and 0.60 for C, respectively. These observations suggest that the discriminative information in a token likelihood sequence is not limited to its global level or overall fluctuation. Instead, the relationship between a salient transition and the subsequent trajectory provides an additional source of information for distinguishing human-written from LLM-generated text.

Motivated by this preliminary analysis, we propose **ExpStabilityDet** (Explicit Stability Detector), a training-free detector that represents a text using both global likelihood characteristics and transition-aware statistics. Given a token log-probability sequence, ExpStabilityDet identifies the dominant transition J and characterizes the variability of the trajectory following that transition through C. These transition-aware features are combined with complementary global likelihood statistics to form the final representation. The key idea is therefore not to introduce another global likelihood statistic, but to explicitly characterize *where a salient likelihood change occurs and how the trajectory behaves afterward*. We further introduce **ExpStabilityDet++**, a sampling-based extension that evaluates the observed text relative to conditionally generated contrast texts, providing a distribution-based variant of the proposed representation. We evaluate the approach across diverse datasets, source models, domains, decoding strategies, text lengths, sampling sizes, and languages, demonstrating the complementary value of transition-aware likelihood characteristics. Our main contributions are:

(i) We identify and empirically validate **transition-aware structure** in token log-probability sequences as a discriminative signal for training-free LLM-generated text detection. In particular, we show that the dominant transition magnitude and the variability of the subsequent trajectory exhibit significant distributional differences between human-written and LLM-generated text.

(ii) We propose **ExpStabilityDet**, a training-free detector that jointly captures global likelihood characteristics, dominant transition magnitude, and post-transition variability. The method provides a simple transition-aware representation without requiring supervised detector training.

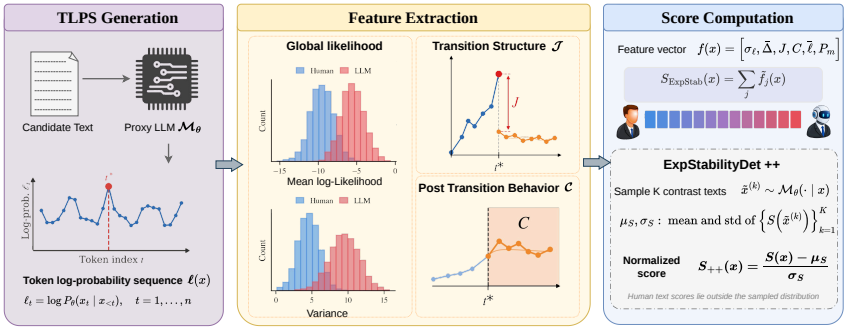


Figure 2: Overview of our proposed ExpStabilityDet framework.

(iii) We extend the proposed approach to **ExpStabilityDet++**, a sampling-based variant that compares the observed text with conditionally generated contrast texts, and conduct extensive evaluations across source models, datasets, decoding strategies, text lengths, multilingual and sampling configurations.

2 Related Work

AI-generated Detection: Existing AI-generated text detection methods can generally be divided into *training-based* and *training-free* approaches. Training-based methods train supervised classifiers to distinguish human-written from LLM-generated text Adam et al. (2026); Liang et al. (2023); Guo et al. (2023); Salminen et al. (2022). While effective on in-domain data, they often generalize poorly to unseen domains, generation models, and out-of-distribution settings Uchendu et al. (2021); Chaka (2024). To address these limitations, training-free methods attempt to exploit intrinsic statistical differences. Early approaches mainly relied on global statistical features such as average log-likelihood and log-rank Solaiman et al. (2019). Later, perturbation-based methods such as DetectGPT Mitchell et al. (2023) introduced the distributional discrepancy paradigm, which compares original text with perturbed variants to measure generation consistency. Subsequent methods such as Fast-DetectGPT Bao et al., Binoculars Hans et al., and DNAGPT Yang et al. improved training-free detection. More recent studies explore sequential modeling of token probability dynamics. More recent studies have explored sequential modeling of token probability dynamics. Lastde Xu et al. interprets the token probability sequence as a time series and captures local temporal fluctuations using diversity entropy. SpecDetect Luo et al. (2026) further extends this analysis into the frequency domain by measuring spectral energy distributions of token probability fluctuations. Despite these advances, existing methods mainly focus on temporal or spectral fluctuations, while transition propagation and stability evolution remain underexplored.

3 Methodology

Given an input text, we obtain its token log-probability sequence, identify the dominant transition and its post-transition variability, and combine these with global likelihood statistics to construct ExpStabilityDet. As illustrated in Figure 2, the detection process follows three main steps

3.1 Token Log-Probability Representation

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ denote an input text containing n tokens, and let $\mathcal{M}_\theta$ denote a proxy language model. For each token position t, the proxy model assigns a conditional probability to the observed token given its preceding context. We represent this probability in logarithmic form as

$$\ell_t = \log P_\theta(x_t \mid x_{<t}), \qquad t = 1, \ldots, n, \tag{1}$$

where $x_{<t} = (x_1, \ldots, x_{t-1})$ denotes the preceding context. The resulting sequence $\mathcal{L}(\mathbf{x}) = (\ell_1, \ell_2, \ldots, \ell_n) \in \mathbb{R}^n$ is referred to as the *token log-probability sequence* (TLPS). The TLPS provides a token-level representation of how well the proxy language model predicts the observed text. We use this sequence to capture both global likelihood characteristics and local changes in predictability.

3.2 FEATURE EXTRACTION

(a) Transition Structure: Existing training-free detectors have demonstrated that token likelihoods contain useful information through global statistics, temporal fluctuations, and frequency-domain characteristics. However, we specifically investigate whether a salient local transition in the likelihood trajectory provides additional discriminative information. To characterize local changes in the TLPS, we first compute the absolute first-order difference between two consecutive token log-probabilities: $\Delta_t = |\ell_{t+1} - \ell_t|, \qquad t = 1, \ldots, n-1$.. The sequence $\{\Delta_t\}_{t=1}^{n-1}$ measures the magnitude of local token-to-token changes in model-assigned log-probability. We then identify the location of the most prominent transition as $t^* = \arg\max_{1 \le t \le n-1} \Delta_t$.. The corresponding *dominant transition magnitude* is explicitly defined

$$J = \Delta_{t^*} = \max_{1 \le t \le n-1} |\ell_{t+1} - \ell_t|. \tag{2}$$

Thus, J measures the magnitude of the largest local change in the token log-probability trajectory, while t^* identifies where that change occurs. In addition to the dominant transition, we calculate the average magnitude of local transitions over the complete sequence:

$$\bar{\Delta} = \frac{1}{n-1} \sum_{t=1}^{n-1} |\ell_{t+1} - \ell_t|. \tag{3}$$

The pair $(J, \bar{\Delta})$ therefore distinguishes the most prominent local transition from the overall level of token-level fluctuation.

(b) Post-Transition Behavior: The dominant transition identifies a salient point in the likelihood trajectory. We next examine whether the behavior of the sequence after this transition provides additional information. Given t^*, we define the post-transition subsequence as $\mathcal{L}_{\text{post}} = (\ell_{t^*+1}, \ell_{t^*+2}, \ldots, \ell_n)$. The mean log-probability of this region is

$$\bar{\ell}_{\text{post}} = \frac{1}{n - t^*} \sum_{t=t^*+1}^{n} \ell_t. \tag{4}$$

We quantify the variability of the post-transition trajectory using its variance:

$$C = \frac{1}{n - t^*} \sum_{t=t^*+1}^{n} \left(\ell_t - \bar{\ell}_{\text{post}}\right)^2. \tag{5}$$

We refer to C as the *post-transition variability*. It quantifies how strongly the likelihood trajectory fluctuates after the dominant transition. Importantly, J and C capture different aspects of the likelihood sequence. J describes the magnitude of the most prominent local change, whereas C describes the subsequent behavior of the trajectory. Consequently, two texts can exhibit similar dominant transition magnitudes while having different post-transition variability.

(c) Global Likelihood Characteristics Although the proposed method focuses on local transition structure, global likelihood statistics provide complementary information about the overall predictability of the observed text. We first compute the mean token log-probability: $\bar{\ell} = \frac{1}{n} \sum_{t=1}^{n} \ell_t$. We also compute the standard deviation of the complete TLPS:

$$\sigma_\ell = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (\ell_t - \bar{\ell})^2}. \tag{6}$$

The mean log-probability captures the overall level of model predictability, whereas σ_ℓ measures the overall variability of token-level likelihoods. Finally, we retain the probability-mass feature formulation. Let τ_{conf} denote a fixed log-probability threshold. We define

$$P_m = \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\left[\ell_t \ge \tau_{\text{conf}}\right], \tag{7}$$

where $\mathbf{1}[\cdot]$ is the indicator function. Following our previous experimental setting, we use $\tau_{\text{conf}} = -1$, corresponding to a token probability of $P_\theta(x_t \mid x_{<t}) \geq e^{-1} \approx 0.37$. Unlike a sequence-dependent percentile threshold, this fixed threshold allows P_m to retain information about the absolute concentration of high-probability tokens.

(d) Transition-Aware Representation The preceding analysis motivates a representation that jointly captures global likelihood characteristics and transition-aware behavior. For an input text $\mathbf{x}$, we construct the feature vector $\mathbf{f}(\mathbf{x}) = [\bar{\ell}, \sigma_\ell, \bar{\Delta}, J, C, P_m]$. The features can be grouped into three complementary components:

1. Global likelihood: $\bar{\ell}$ and σ_ℓ characterize the overall predictability and variability of the text.

2. Transition structure: $\bar{\Delta}$ and J characterize token-level changes and identify the most prominent local transition.

3. Post-transition behavior: C characterizes the variability of the likelihood trajectory following the dominant transition, while P_m captures the concentration of tokens above a fixed likelihood threshold.

The central component of our representation is the pair (J, C). Rather than treating each token likelihood independently, this pair explicitly represents both the magnitude of a salient transition and the behavior that follows it.

3.3 EXPSTABILITYDET: SCORE COMPUTATION

Based on the transition-aware representation, we propose **ExpStabilityDet** (*Explicit Stability Detector*), a training-free detector for distinguishing human-written and LLM-generated text. Given an input text $\mathbf{x}$, the detector follows three steps.

Step 1: TLPS generation. The input text is passed through a proxy language model $\mathcal{M}_\theta$ to obtain the token log-probability sequence $\mathcal{L}(\mathbf{x})$.

Step 2: Feature extraction. The TLPS is used to calculate the global, transition, and post-transition features.

Step 3: Score computation. To make the features comparable, each feature is normalized using a held-out calibration set. Let $\tilde{f}_j(\mathbf{x})$ denote the normalized value of the j-th feature. The detector score is computed as a calibrated combination of these normalized features:

$$S_{\text{ExpStab}}(\mathbf{x}) = \sum_{j=1}^{6} w_j \tilde{f}_j(\mathbf{x}), \tag{8}$$

where w_j denotes the calibrated contribution of the corresponding feature. The calibration procedure is performed independently of the evaluation test sets. The calibration set is used only to normalize the feature ranges and determine the score parameters. The final decision is obtained by comparing the detector score with a threshold τ:

$$\hat{y}(\mathbf{x}) = \texttt{human}, \quad S_{\text{ExpStab}}(\mathbf{x}) \geq \tau, ; \qquad \hat{y}(\mathbf{x}) = \texttt{LLM}, \quad S_{\text{ExpStab}}(\mathbf{x}) < \tau. \tag{9}$$

3.4 SAMPLING-BASED EXPSTABILITYDET++

The single-pass detector evaluates a text using one observed TLPS. However, the score of an individual text can depend on the particular token sequence being evaluated. To obtain a distribution-based assessment, we extend ExpStabilityDet using conditionally generated contrast texts. Given an input text $\mathbf{x}$, we generate K contrast texts $\{\tilde{\mathbf{x}}^{(1)}, \tilde{\mathbf{x}}^{(2)}, \ldots, \tilde{\mathbf{x}}^{(K)}\} \sim \mathcal{M}_\theta(\cdot \mid \mathbf{x})$, where each contrast text is generated conditioned on the input context. For the K contrast texts, we compute the mean μ_S and standard deviation σ_S of their detector scores. The final distribution-based score is then defined as

$$S^{++}_{\text{ExpStab}}(x) = \frac{S_{\text{ExpStab}}(x) - \mu_S}{\sigma_S + \epsilon}, \tag{10}$$

where $\epsilon = 10^{-8}$ ensures numerical stability.

Table 1: AUROC averaged over XSum, WritingPrompts, and Reddit. Best: **bold**; second: underlined.

Method	BLOOM-7B	FALCON-7B	GEMMA-7B	QWEN2.5-3B	GPTJ-6B	GPTNEO-2	LLAMA3-8B	LLAMA2-13B	OPT-13B	TinyLlama	PHI4	Avg.
Sample-based Methods												
Likelihood	0.6180	0.6415	0.6990	0.5787	0.8707	0.6708	0.6574	0.6861	0.6881	0.6739	0.7394	0.6840
Logrank	0.6748	0.7168	0.7215	0.6241	0.9010	0.7117	0.7030	0.7267	0.7304	0.7230	0.7803	0.7285
Entropy	0.6084	0.5055	0.4803	0.6530	0.4998	0.5866	0.4913	0.4518	0.5310	0.5454	0.4659	0.5290
DetectLRR	0.7954	0.7990	0.7354	0.7212	0.9257	0.7927	0.7853	0.7888	0.8031	0.8111	0.8378	0.7996
Lastde	0.8545	0.8278	0.7794	0.8532	0.9749	0.8681	0.7787	0.7891	0.8904	0.8912	0.8591	0.8515
SpecDetect	0.8780	0.8394	0.8168	**0.8982**	**0.9799**	0.9170	0.8050	0.8089	0.9005	0.9118	0.9085	0.8785
ExpStabilityDet	**0.9202**	**0.8714**	0.8062	0.8970	0.9783	**0.9326**	**0.8187**	**0.8113**	**0.9046**	**0.9163**	0.9076	**0.8877**
Distribution-based Methods												
DetectGPT	0.5468	0.5825	0.6819	0.6431	0.6656	0.6086	0.5600	0.5899	0.5933	0.5972	0.5950	0.6058
DetectNPR	0.6238	0.6460	0.6022	0.6887	0.7390	0.6840	0.6037	0.6076	0.6704	0.6637	0.6645	0.6539
DNA-GPT	0.7015	0.7244	0.6868	0.7292	0.7804	0.7514	0.6717	0.6840	0.7629	0.7395	0.7129	0.7223
Fast-DetectGPT	0.6466	0.6914	0.7002	0.7002	0.7932	0.7206	0.6744	0.6512	0.7051	0.7048	0.6779	0.6969
Lastde++	0.8613	0.8245	0.7791	0.8570	0.9176	0.8820	0.7760	0.7860	0.8090	0.8281	0.8632	0.8349
SpecDetect++	**0.9245**	0.8731	0.8008	0.9076	0.9290	**0.9483**	0.8464	0.8034	0.9313	0.9075	0.8973	0.8881
ExpStabilityDet++	0.9178	0.8796	**0.8689**	**0.9203**	0.9759	0.9429	**0.8665**	0.8696	**0.9446**	**0.9526**	**0.9346**	**0.9157**

Table 2: Performance comparison of training-free detectors on the WMT16 (German) and Zhihu (Chinese) datasets. Higher scores indicate better detection performance.

Dataset	Lastde	SpecDetect	ExpStabilityDet	Lastde++	SpecDetect++	ExpStabilityDet++
WMT16	0.6368	0.7418	**0.8128**	0.6707	0.6893	**0.7373**
Zhihu	0.9474	0.9558	**0.9658**	0.8827	0.9257	**0.9460**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. Our experiments are conducted on five different domain datasets and languages as mentioned SpecDetect Luo et al. (2026). We primarily report detection performance on three widely used datasets: XSum Narayan et al. (2018) for BBC news documents, WritingPrompts Fan et al. (2018) for long-form story, and Reddit ELI5 Fan et al. (2019) containing Q&A discussions from domains such as biology, physics, chemistry, economics, law, and technology. To further analyze cross-lingual robustness, we additionally evaluate on WMT16 Bojar et al. (2016), which contains English–German translation pairs, and a Chinese dataset collected from the Zhihu platform covering topics including food, history, and economics. Each dataset consists of 150 human-written samples. For every human-written instance, the first 30 tokens are used as a prompt to the source language model to generate a continuation, producing the corresponding AI-generated text. Additional details on dataset preparation and prompt construction are provided in Appendix B.1.

Source models & proxy models. We evaluate our framework on a diverse set of open-source and closed-source LLMs, including 11 open-source models (1.1B to 13B) and three closed-source systems: GPT-4-Turbo, GPT-5, and Claude-3-haiku, following prior works Xu et al.; Luo et al. (2026). To evaluate performance on more recent architectures, we introduce four new LLM datasets: Qwen2.5-3B, Llama3-8B, TinyLlama, and Phi-4. For multilingual robustness experiments on German and Chinese datasets, we include specialized models such as mGPT, Qwen1.5, and Yi-1.5. The primary detection results are reported using the 11 open-source models listed in Table 1. Unless otherwise specified, GPT-J-6B is adopted as the default proxy model for all black-box experiments 7. It is utilized throughout the entire detection pipeline, including rewriting, sampling, and scoring stages. Complete specifications for all source models are provided in Appendix 8.2.

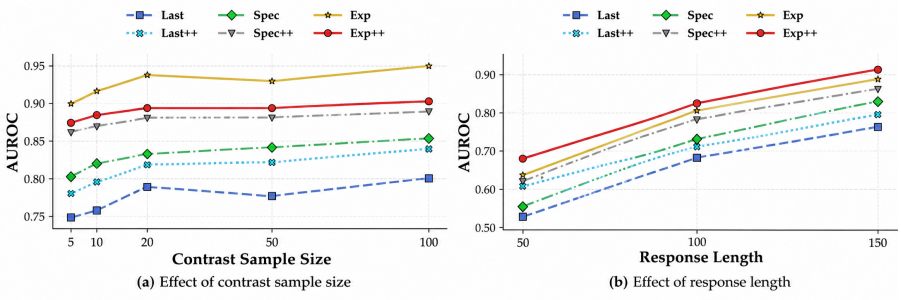
Baselines. We compare against twelve representative training-free baselines in two categories: The sample-based baselines include Log-Likelihood Gehrmann et al. (2019), LogRank Gehrmann et al. (2019), Entropy Gehrmann et al. (2019), and DetectLRR Hans et al., Lastde Xu et al. and SpecDetect Luo et al. (2026). The distribution-based baselines consist of DetectGPT Mitchell et al. (2023), DetectNPR Hans et al., DNA-GPT Yang et al., and Fast-DetectGPT Bao et al. , Lastde++ Xu et al. Luo et al. (2026) and SpecDetect++. Detailed descriptions of these baselines are available in Appendix 8.3. For evaluation, we follow previous work and report the Area Under the ROC Curve (AUROC) as the primary metric. All experiments are conducted on a RTX A5000 with 24GB of memory.

4.2 MAIN RESULTS

We primarily evaluate ExpStabilityDet in the black-box setting (see Appendix 7), which closely reflects realistic deployment scenarios. Table 1 summarizes the average AUROC across XSum, WritingPrompts, and Reddit, while detailed per-dataset results are provided in Appendix Tables 16,

Table 3: Statistical comparison of likelihood-profile between human-written and LLM-generated text.

Feature	Human Median	LLM Median	p (Holm)	Cliff's δ
Mean log-probability $\bar{\ell}$	-2.735	-2.410	$< 10^{-5}$	-0.49
Overall likelihood variability σ_ℓ	2.532	1.952	$< 10^{-5}$	0.88
Mean transition magnitude $\bar{\Delta}$	2.695	2.169	$< 10^{-5}$	0.80
Dominant transition magnitude J	11.029	7.743	$< 10^{-5}$	0.77
Post-transition variance C	6.345	3.715	$< 10^{-5}$	0.70
Probability-mass concentration P_m	0.325	0.312	0.198	0.05



(a) Effect of contrast sample size **(b)** Effect of response length

Figure 3: Average AUROC performance across three datasets under different text lengths and different numbers of contrast samples.

17, and 18. In the sample-based setting, ExpStabilityDet achieves an average AUROC of 0.8877 across the evaluated source models, compared with 0.8785 for SpecDetect and 0.8515 for Lastde. ExpStabilityDet performs particularly well for BLOOM-7B, FALCON-7B, GPTNeo-2, LLaMA3-8B, LLaMA2-13B, OPT-13B, and TinyLlama. These results indicate that incorporating transition structure and post-transition variability into the token log-probability representation provides a useful signal beyond global likelihood-based statistics.

We further evaluate the sampling-based extension, ExpStabilityDet++, in the distribution-based setting. ExpStabilityDet++ obtains an average AUROC of 0.9157, compared with 0.8881 for SpecDetect++ and 0.8349 for Lastde++. The improvement is observed across several source models, although the relative performance varies across generators. This variation suggests that the effectiveness of the proposed representation depends on the characteristics of the source model and motivates the component-level analysis presented in Section 4.3 and Table 1. Overall, these results support the central premise of ExpStabilityDet: local transitions in token log-probability sequences, together with the variability following a dominant transition, provide complementary information for distinguishing human-written and LLM-generated text. Importantly, ExpStabilityDet does not require model training or access to the generator model. Instead, it derives the detection signal directly from the token-level likelihood behavior of a proxy language model. Detailed white-box experimental results are presented in Appendix Section 10.3 and Table 10. Results for closed-source LLMs, including GPT-5, and additional benchmark evaluations against existing methods on HC3, RAID, and M4 are provided in Appendices 12.1 and 12.2.

4.3 Statistical Validation

We also examine whether the likelihood-profile features defined in Section 3 exhibit systematic differences between human-written and LLM-generated text. Table 3 reports the median values for both classes together with Holm-corrected Mann–Whitney U test p-values and Cliff's δ effect sizes. Mean log-probability, overall likelihood variability, mean transition magnitude, dominant transition magnitude, and post-transition variance all exhibit statistically significant differences between the two classes ($p < 10^{-5}$). The effect sizes are particularly large for overall likelihood variability ($\delta = 0.88$), mean transition magnitude ($\delta = 0.80$), dominant transition magnitude ($\delta = 0.77$), and post-transition variance ($\delta = 0.70$). These results indicate that both global likelihood characteristics and transition-aware statistics contain discriminative information.

Component ablation. We further analyze the contribution of individual features and feature groups in Appendix 9. The complete six-feature representation achieves a mean AUROC of 0.9820, compared

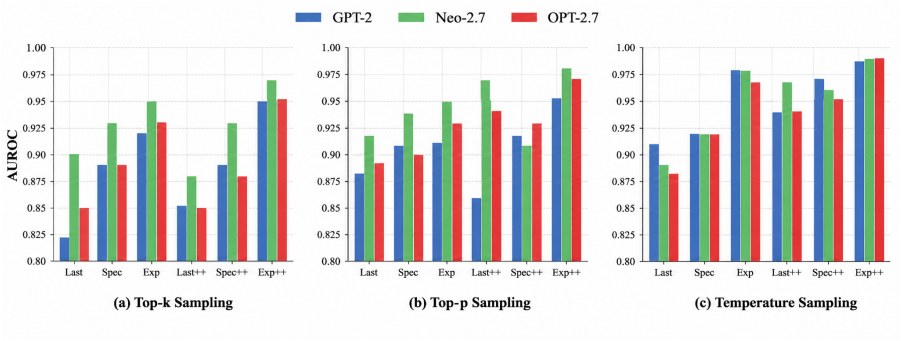
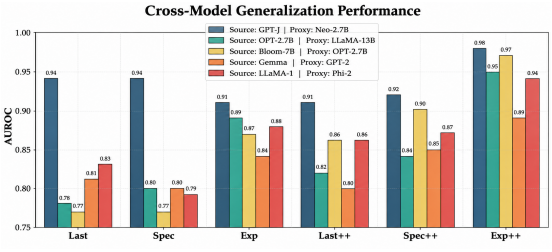
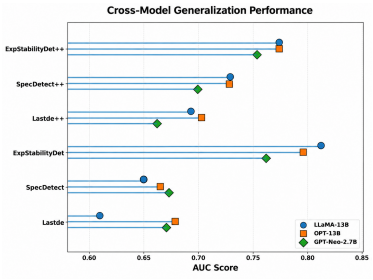


Figure 4: Avg. AUROC across three datasets under top-k, top-p, and temperature sampling for three source models. Detailed results in Table 13



((a)) Different source–proxy model pairs.



((b)) Impact of paraphrasing under black-box settings.

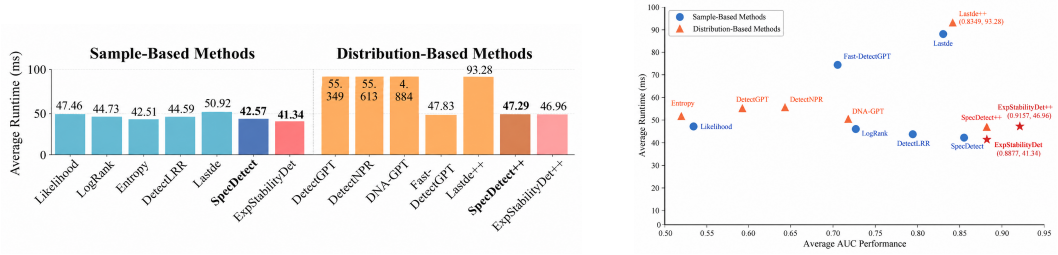
Figure 5: (a) Detection performance across different source–proxy model pairs on the XSum dataset. (b) Detection robustness under paraphrasing attacks across different source models.

with 0.9799 for the strongest individual feature (σ_ℓ). The leave-one-feature-out results show that removing any component reduces the mean AUROC relative to the full representation, with the largest decrease observed when removing probability-mass concentration P_m. The group-wise analysis further shows that transition-aware statistics provide strong standalone discrimination, while the full combination of global, transition, post-transition, and probability-mass features gives the highest performance among the evaluated configurations.

4.4 Robustness Analysis

Response Lengths. To evaluate the robustness of ExpStabilityDet to varying response lengths, we truncate GPT-4-Turbo generations from the XSum dataset to {50, 100, 150} words. Figure 3 presents the corresponding detection performance, with detailed results provided in Appendix 11.1. Overall, detection performance improves as the response length increases, indicating that longer responses provide more token-level likelihood information for discrimination. Its sampling-based extension, ExpStabilityDet++, further improves the detection signal as the available response length increases. These results demonstrate that the transition- and post-transition stability features remain informative across different response lengths.

Sample Size. Following the setting of prior work Luo et al. (2026), we evaluate the effect of different numbers of contrastive samples on detection performance. Figure 3 reports the results for contrast sample sizes of {5, 10, 20, 50, 100}. Overall, the detection performance of the sampling-based methods improves as the number of contrastive samples increases, although the magnitude of improvement varies across methods. ExpStabilityDet++ maintains strong performance across the evaluated sample sizes, indicating that the proposed stability-aware representation can effectively exploit additional contrastive information. The relatively small performance variation across larger sample sizes also suggests that the method does not strongly depend on a specific sample size. Per-dataset results are provided in Appendix Section 11.2.



((a)) Average runtime comparison.

((b)) AUROC–runtime trade-off.

Figure 6: (a) Average inference runtime across sample-based and distribution-based methods. (b) Relationship between average AUROC and runtime.

Decoding Strategies. We further evaluate the robustness of training-free detectors under multiple decoding strategies, including top-k, top-p, and temperature sampling, which control generation diversity in different ways.Figure 4 summarizes the results, with detailed results provided in Appendix 11.3. The results show that decoding strategy can substantially affect detection performance across source models and detector variants. ExpStabilityDet maintains strong performance across the evaluated decoding strategies, while its sampling-based extension, ExpStabilityDet++, remains effective under different generation settings. These results suggest that the transition- and post-transition stability signals remain informative under variations in generation diversity and sampling behavior.

Generalization to Different Source–Proxy Model Pairs. The choice of proxy model can affect black-box detection performance. To evaluate the robustness of ExpStabilityDet under different source–proxy configurations, we conduct experiments on XSum using five source–proxy model pairs, as shown in Figure 5(a). Performance varies across pairs, with larger capacity gaps generally leading to lower detection performance. In particular, pairing $\sim$7B source models with 2.7B proxy models results in weaker performance in several settings, highlighting the importance of source–proxy compatibility.

Paraphrasing Attack. We further evaluate ExpStabilityDet under paraphrasing attacks, where LLM-generated texts are produced by paraphrasing human-written text. Figure 5(b) shows performance across different source models. Paraphrasing substantially affects training-free detectors, while ExpStabilityDet maintains competitive performance, suggesting that its transition- and post-transition stability features remain informative beyond surface-level lexical patterns.

Non-English Scenarios. To evaluate cross-lingual generalization, we conduct experiments on German (WMT16) and Chinese (Zhihu), with results shown in Table 2. ExpStabilityDet maintains strong performance across both languages, achieving AUROC scores of 0.8128 on WMT16 and 0.9658 on Zhihu. In comparison, SpecDetect achieves 0.7418 and 0.9558, respectively. The sampling-based extension, ExpStabilityDet++, further achieves AUROC scores of 0.7373 on WMT16 and 0.9460 on Zhihu. These results indicate that the transition- and post-transition stability representation remains informative across languages with different linguistic characteristics.

Runtime and Performance–Efficiency Analysis Figure 6 compares detection performance with inference cost. ExpStabilityDet achieves 0.8877 AUROC with only 41.34 ms runtime, while ExpStabilityDet++ improves the AUROC to 0.9157 with 46.96 ms runtime. Thus, both variants provide a favorable performance-efficiency trade-off compared with the evaluated baselines.

5 Conclusion

We proposed ExpStabilityDet, a training-free detector that combines global likelihood statistics with dominant transition magnitude and post-transition variability, along with ExpStabilityDet++ for distribution-based detection. Across diverse datasets, source models, decoding strategies, response lengths, and languages, ExpStabilityDet achieves an average AUROC of 0.8877, while ExpStabilityDet++ achieves 0.9157. Our results demonstrate that transition-aware likelihood statistics provide a useful perspective for training-free LLM-generated text detection. We also find that proxy-model selection affects black-box performance, motivating future work on adaptive proxy selection and more robust transition-aware representations.

AI USE STATEMENT

Generative AI tools were only used to assist with language editing, grammar correction, and refinement of the manuscript text. All experimental designs, implementations, analyses, results, and scientific claims were reviewed and verified by the authors. The authors take full responsibility for the final content of the manuscript.

ETHICS STATEMENT

This work uses publicly available datasets and does not involve human subjects, collection of personally identifiable information, or direct interaction with human participants. The proposed method is intended for research on the detection of LLM-generated text. We acknowledge that automated detection systems can produce false positives and false negatives and should therefore not be treated as definitive evidence of authorship without appropriate human review.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed feature construction, detector scoring procedure, sampling-based extension, datasets, source and proxy models, baselines, and evaluation protocol in the main paper and supplementary material. The experiments use publicly available datasets and language models where applicable. The experimental setup, including the proxy-model configuration, contrast-sample size, evaluation metrics, and robustness settings, is described to facilitate reproduction of the reported results. Source code and additional experimental configurations can be released as supplementary material.

REFERENCES

George Alexandru Adam, Alexander Cui, Edwin Thomas, Emily Napier, Nazar Shmatko, Jacob Schnell, Jacob Junqi Tian, Alekhya Dronavalli, Edward Tian, and Dongwon Lee. Gptzero: Robust detection of llm-generated texts. *arXiv preprint arXiv:2602.13042*, 2026.

Aaron Adcock, Aayushi Srivastava, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pande, Abhinav Pandey, Abhinav Sharma, Abhishek Kadian, Abhishek Kumawat, Adam Kelsey, et al. The llama 4 herd: Architecture, training, evaluation, and deployment notes. *arXiv preprint arXiv:2601.11659*, 2026.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detec-tion of machine-generated text via condi-tional probability curvature.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 131–198, 2016.

Chaka Chaka. Reviewing the performance of ai detection tools in differentiating between ai-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning and Teaching*, 7(1), 2024.

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

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3558–3567, 2019.

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

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

A Hans, A Schwarzschild, V Cherepanova, H Kazemi, A Saha, M Goldblum, J Geiping, and T Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text, 2024. *URL: https://arxiv. org/abs/2401.12070*.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7), 2023.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. Coco: Coherence-enhanced machine-generated text detection under data limitation with contrastive learning. *arXiv preprint arXiv:2212.10341*, 2022.

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models. *arXiv preprint arXiv:2304.07666*, 2023.

Haitong Luo, Weiyao Zhang, Suhang Wang, Wenji Zou, Chungang Lin, Xuying Meng, and Yujun Zhang. Specdetect: Simple, fast, and training-free detection of llm-generated text via spectral analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32356–32364, 2026.

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

Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1797–1807, 2018.

Seyeon Park and Xiaoli Nan. Generative ai and misinformation: a scoping review of the role of generative ai in the generation, detection, mitigation, and impact of misinformation. *AI & SOCIETY*, pp. 1–15, 2025.

Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64:102771, 2022.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

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

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

Teo Susnjak and Timothy R McIntosh. Chatgpt: The end of online exam integrity? *Education Sciences*, 14(6):656, 2024.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. Turingbench: A benchmark environment for turing test in the age of neural text generation. *arXiv preprint arXiv:2109.13296*, 2021.

Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. Training-free llm-generated text detection by mining token probability sequences.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Ruth Petzold, William Yang Wang, and Haifeng Chen. Dna-gpt: Divergent ngram analysis for training-free detection of gpt-generated text.

Xiaowei Zhu, Yubing Ren, Fang Fang, Qingfeng Tan, Shi Wang, and Yanan Cao. Dna-detectllm: Unveiling ai-generated text via a dna-inspired mutation-repair paradigm. *arXiv preprint arXiv:2509.15550*, 2025.

APPENDIX

LIMITATIONS

Our study has several limitations. First, black-box detection performance depends on the choice of proxy language model, and substantial differences between source and proxy model capacities can reduce detection performance. Second, although we evaluate multiple datasets, source models, languages, decoding strategies, and paraphrasing settings, the evaluation does not cover all possible generators and text domains. Third, the sampling-based ExpStabilityDet++ requires multiple contrast generations and therefore introduces additional computational cost compared with the single-pass detector.

6 TASK AND SCENARIO DEFINITION FOR TRAINING-FREE DETECTION

Task Definition. Following Lastde Xu et al. and SpecDetect Luo et al. (2026), we define the training-free detection problem as follows. Given a candidate text $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ of n tokens, determine whether $\mathbf{x}$ was generated by a large language model, without any supervised training of the detector. Any available open-source language model may be used as a proxy model M_θ for inference and scoring. This definition allows the detector to operate in a fully unsupervised manner, making no assumptions about the identity or architecture of the source model that produced $\mathbf{x}$.

White-Box Setting. In the white-box setting, the proxy model M_θ used for detection is identical to the source model M_ϕ that generated the candidate text. This provides the strongest possible discriminative signal, since the log-probability sequence $\ell(\mathbf{x})$ is computed under the exact generation distribution. The white-box setting corresponds to the idealized scenario of known-source detection and serves as an upper-bound evaluation for training-free methods.

Black-Box Setting. In the black-box setting, the identity of the source model M_ϕ is unknown, and the detector must rely on a separate open-source proxy model $M_\theta \neq M_\phi$ for all inference and scoring operations. This is the more realistic and practically relevant setting, as it corresponds to detecting text from closed-source commercial APIs (e.g., GPT-4-Turbo, Claude-3-haiku) using only publicly available proxy models. Unless otherwise specified, GPT-J-6B is used as the default proxy model in all black-box experiments throughout this work.

7 ADDITIONAL DETAILS OF METHODOLOGY

7.1 FIXED THRESHOLD FOR PROBABILITY-MASS CONCENTRATION

The probability-mass concentration feature P_m (Equation 12 of the main paper) requires a sequence-independent log-probability threshold τ_{conf}. We set $\tau_{\text{conf}} = -1$, corresponding to

$$P_\theta(x_t \mid x_{<t}) \geq e^{-1} \approx 0.37. \tag{11}$$

Thus, P_m measures the fraction of tokens whose conditional probability under the proxy language model exceeds approximately 0.37. We use a fixed threshold rather than a sequence-dependent percentile so that P_m retains information about the absolute concentration of high-probability tokens across different texts.

7.2 CALIBRATION OF EXPSTABILITYDET

ExpStabilityDet combines six complementary features: $\bar{\ell}$, σ_ℓ, $\bar{\Delta}$, J, C, and P_m. As defined in Equation 14 of the main paper, each feature is first normalized using a held-out calibration set and then combined using calibrated weights:

$$S_{\mathrm{ExpStab}}(x) = \sum_{j=1}^{6} w_j \widetilde{f}_j(x), \tag{12}$$

where $\widetilde{f}_j(x)$ denotes the normalized value of the j-th feature and w_j is its calibrated contribution.

The calibration set is kept separate from all evaluation test sets. It is used only to determine the feature normalization parameters and the score parameters. This ensures that the evaluation datasets are not used for calibration.

7.3 DECISION THRESHOLD SELECTION

The decision threshold τ in Equation 15 of the main paper is selected using the held-out calibration set. Specifically, we compute the distribution of S_{ExpStab} for the human-written calibration samples and select the threshold corresponding to the target false-positive-rate constraint. The resulting threshold is then fixed and applied to the evaluation data.

Importantly, the primary evaluation metric in this work is AUROC, which is threshold-independent. Therefore, the threshold selection does not affect the reported AUROC values. Threshold-dependent metrics, when reported, use the threshold determined from the calibration set.

7.4 SAMPLING-BASED EXPSTABILITYDET++

ExpStabilityDet++ extends the single-pass ExpStabilityDet detector by evaluating the observed text relative to conditionally generated contrast texts. Given an input text x, we generate K contrast texts

$$\left\{\widetilde{x}^{(1)}, \widetilde{x}^{(2)}, \ldots, \widetilde{x}^{(K)}\right\} \sim M_\theta(\cdot \mid x), \tag{13}$$

where each contrast text is generated conditioned on the input context. For every contrast text, we compute the corresponding ExpStabilityDet score:

$$S_k = S_{\mathrm{ExpStab}}\left(\widetilde{x}^{(k)}\right). \tag{14}$$

The mean and standard deviation of the contrast-score distribution are then calculated as

$$\mu_S = \frac{1}{K}\sum_{k=1}^{K} S_k, \tag{15}$$

and

$$\sigma_S = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(S_k - \mu_S)^2}. \tag{16}$$

The final distribution-normalized score is

$$S^{++}_{\mathrm{ExpStab}}(x) = \frac{S_{\mathrm{ExpStab}}(x) - \mu_S}{\sigma_S + \epsilon}, \tag{17}$$

where $\epsilon = 10^{-8}$ is used for numerical stability. Unless otherwise specified, the main experiments use $K = 100$ contrast samples.

8 Detail About The Setting And Baselines

8.1 Dataset Details

Our experiments are conducted on six benchmark datasets spanning multiple domains and languages. Table 4 summarises the key properties of each dataset.

Table 4: Summary of datasets used in experiments.

Dataset	Domain	Language	Samples	Setting
XSum	News summarization	English	150	White-box & Black-box
WritingPrompts	Story generation	English	150	White-box & Black-box
Reddit ELI5	Explanatory Q&A	English	150	Black-box
WMT16	Translation	German	150	Cross-lingual
Zhihu	Q&A (Economy/Food/History)	Chinese	150	Cross-lingual

For each dataset, we filter out samples with fewer than 150 words and randomly sample 150 human-written instances. For each human-written instance, the first 30 tokens are used as a prompt to the source language model to generate a continuation of approximately 200 words, producing the corresponding LLM-generated text. Each finalized dataset therefore consists of 150 negative samples (human-written) and 150 positive samples (LLM-generated), with each pair sharing the same initial prefix tokens, ensuring matched text length and consistent evaluation conditions.

For source models requiring system prompts or API calls, including GPT-4-Turbo, GPT-5, and Claude-3-haiku, we use the prompt templates shown in Table 5.

Table 5: Prompt templates used for closed-source API models.

Dataset	System Prompt	User Prompt
XSum	You are a professional news writer.	Please write an article with about 200 words starting exactly with: *[Prefix]*
WritingPrompts	You are a professional fiction writer.	Please write a story with about 200 words starting exactly with: *[Prefix]*
Reddit ELI5	You are a helpful assistant that answers the question provided.	Please continue the answer in 200 words starting exactly with: *[Prefix]*

8.2 ALL LARGE LANGUAGE MODELS USED

Table 6 lists all source models used to generate LLM-generated text in our experiments, including both open-source and closed-source systems.

Table 6: Specification of all source models used in experiments.

Model	Source / Repository	Parameters	Type
Qwen2.5-3B	`Qwen/Qwen2.5-3B`	3B	Open-source
GPT-Neo-2.7B	`EleutherAI/gpt-neo-2.7B`	2.7B	Open-source
TinyLlama	`TinyLlama/TinyLlama-1.1B-Chat-v1.0`	1.1B	Open-source
GPT-J-6B	`EleutherAI/gpt-j-6b`	6B	Open-source
BLOOM-7B1	`bigscience/bloom-7b1`	7.1B	Open-source
Falcon-7B	`tiiuae/falcon-7b`	7B	Open-source
Gemma-7B	`google/gemma-7b`	7B	Open-source
LLaMA2-13B	`meta-llama/Llama-2-13b`	13B	Open-source
LLaMA3-8B	`meta-llama/Meta-Llama-3-8B`	8B	Open-source
OPT-13B	`facebook/opt-13b`	13B	Open-source
Phi-4-14B	`microsoft/phi-4`	14B	Open-source
mGPT	`ai-forever/mGPT`	1.3B	Open-source (multilingual)
Qwen1.5-7B	`Qwen/Qwen1.5-7B`	7B	Open-source (multilingual)
Yi-1.5-6B	`01-ai/Yi-1.5-6B`	6B	Open-source (multilingual)
GPT-4-Turbo	OpenAI API	N/A	Closed-source
GPT-5	OpenAI API	N/A	Closed-source
Claude-3-haiku	Anthropic API	N/A	Closed-source

Unless otherwise specified, GPT-J-6B is adopted as the default proxy model in all black-box experiments. For multilingual experiments on German and Chinese datasets, mGPT and Yi-1.5-6B are used as proxy models respectively, following the setup of Lastde Xu et al.. All open-source models

use half-precision (float16) except GPT-2-XL, GPT-Neo-2.7B, OPT-2.7B, and mGPT, which use full precision (float32) due to numerical stability requirements.

8.3 Baselines details

8.3.1 Sample-based Methods

We compare our methods against ten representative training-free baselines divided into two categories.

8.3.2 Sample-Based Methods

Log-Likelihood Solaiman et al. (2019): The average log-probability of all tokens in the candidate text, $\bar{\ell} = \frac{1}{n}\sum_{i=1}^{n} \ell_i$, is used as the detection score. LLM-generated text typically exhibits a higher (less negative) mean log-likelihood.

LogRank Solaiman et al. (2019): The average log-rank of each token, where rank is computed by sorting the full vocabulary in descending order of probability at each position. LLM-generated text tends to select higher-ranked (lower log-rank) tokens.

Entropy Gehrmann et al. (2019): The average Shannon entropy of the next-token distribution at each position, $\bar{H} = \frac{1}{n}\sum_{i=1}^{n} H_i$. LLM-generated text tends to exhibit lower entropy at each generation step.

DetectLRR Su et al. (2023): The ratio of average log-likelihood to average log-rank. This ratio normalises the absolute probability level against the relative token rank, providing a more discriminative signal than either statistic alone.

Lastde Xu et al.: Treats the token probability sequence as a time series and computes the Multiscale Diversity Entropy (MDE) of the sequence. The detection score is the ratio of log-likelihood to aggregated MDE, capturing both global and local temporal dynamics. Default hyperparameters: $s = 3$, $\varepsilon = 10 \times n$, $\tau' = 5$.

SpecDetect Luo et al. (2026): Applies the Discrete Fourier Transform (DFT) to the zero-mean log-probability sequence and uses the total spectral energy $E_{\mathrm{DFT}} = \sum_k |X_k|^2$ as the detection score, treating human text as having higher generative vitality in the frequency domain.

8.3.3 Distribution-based Methods.

9 Extensive Ablation Study results

9.1 Effect of the Number of Contrast Samples

We additionally evaluate the sensitivity of the sampling-based (++) detectors to the number of conditional contrast samples K. We consider $K \in \{5, 10, 20, 50, 100\}$, with the results reported in Table 7. For ExpStabilityDet++, performance improves from 0.9778 at $K = 5$ to 0.9890 at $K = 10$, 0.9926 at $K = 20$, and 0.9942 at $K = 50$. Increasing K further to 100 results in a slight decrease to 0.9932. Thus, the performance largely saturates around $K = 20$–50, with $K = 50$ providing the highest AUROC among the evaluated settings. In contrast, Lastde++ and SpecDetect++ exhibit smaller and less monotonic variations across different sample sizes. These results indicate that increasing the number of contrast samples beyond 50 does not provide a consistent performance benefit for ExpStabilityDet++.

Computational overhead. The additional computation in the ++ variants consists of conditional contrast-sample generation and the subsequent contrastive normalization. Once the contrast samples have been scored, the normalization itself only requires computing the mean and standard deviation of the K contrast scores followed by score normalization. In our efficiency measurement, this contrastive normalization procedure for $K = 100$ introduces an additional 10.44 ms per text, corresponding to approximately 13.3% of the cost of a single language-model forward pass. Thus, the normalization stage introduces relatively small computational overhead compared with the language-model computation required to obtain the contrast samples.

Table 7: Effect of the number of conditional contrast samples K on the AUROC of sampling-based detectors.

K	Lastde++	SpecDetect++	ExpStabilityDet++
5	0.9551	0.9562	0.9778
10	0.9505	0.9741	0.9890
20	0.9502	0.9698	0.9926
50	0.9630	0.9716	**0.9942**
100	0.9539	0.9709	0.9932

Table 8: Representative source–proxy model pairs and their corresponding AUROC. The parameter gap denotes the absolute difference in parameter count between the source and proxy models.

Source LLM	Proxy LLM	Parameter Gap	AUROC
Llama-3.1-8B	Llama-3.1-8B	0B	0.896
Llama-3.1-8B	Llama-3.2-3B	5B	**0.923**
Llama-3.1-8B	Llama-3.2-1B	7B	0.834
Gemma-9B	Gemma-9B	0B	0.882
Gemma-9B	Gemma-2B	7B	0.823

10 Additional Details of Results Detection Methods

DetectGPT Mitchell et al. (2023): Generates 100 perturbations of the candidate text using T5-3B LLM model and measures the discrepancy between the original text's log-likelihood and the mean log-likelihood of its perturbations. Texts whose log-likelihood is locally maximal are classified as LLM-generated.

10.1 Additional Source–Proxy Analysis

To further quantify the effect of source–proxy model selection, we summarize representative source–proxy pairs in Table 8. The results illustrate the variation in AUROC across different proxy choices for the same source LLM.

These results show that detector performance varies across proxy choices. For Llama-3.1-8B, the Llama-3.2-3B proxy achieves an AUROC of 0.923, compared with 0.896 using the same model as the proxy and 0.834 using Llama-3.2-1B. Similarly, for Gemma-9B, the same-model proxy achieves 0.882, while the Gemma-2B proxy achieves 0.823. These results suggest that proxy selection is not determined solely by parameter count; model family and capacity may jointly affect the resulting likelihood estimates.

DetectNPR Su et al. (2023): A normalised variant of DetectGPT that uses the normalized perturbation discrepancy in log-rank space rather than log-likelihood space, improving robustness to the absolute probability level.

DNA-GPT Yang et al.: Generates contrast samples by truncating the candidate text at a fixed rate and rewriting the truncated portion using the proxy model. The log-likelihood divergence between original and rewritten segments serves as the detection score. Default settings: truncation rate $\tau = 0.5$, 10 rewrites.

Fast-DetectGPT Bao et al.: Replaces DetectGPT's expensive perturbation sampling with conditional probability curvature computed via fast sampling. We use GPT-J as both the sampling and scoring model for fair comparison (deviating from the original setting which uses GPT-J for sampling and Neo-2.7B for scoring). Sampling size: 10,000.

Lastde++ Xu et al.: The sampling-based extension of Lastde, which normalises the Lastde score via a z-score relative to the distribution of Lastde scores over conditionally sampled texts. Default settings: $s = 4$, $\varepsilon = 8 \times n$, $\tau' = 15$, 100 samples.

Table 9: Component-wise ablation of ExpStabilityDet on two Reddit source-model settings (Reddit-GPT2-XL and Reddit-GPTJ-6B). Global denotes $\{\mathrm{mean}, \mathrm{std}\}$, Transition denotes $\{\mathrm{diff}, J\}$, and Post-transition denotes $\{C\}$. Mean, Worst, and Std. denote the mean, worst-case, and standard deviation of AUROC, respectively, across the two settings. Δ denotes the difference in mean AUROC relative to the full configuration.

Configuration	Features	Mean	Worst	Std.	Δ
Individual features					
Mean only	mean	0.7786	0.6688	0.1098	-0.2034
Std. only	std	0.9799	0.9670	0.0129	-0.0021
Diff. only	diff	0.9425	0.9023	0.0402	-0.0395
J only	J	0.9413	0.8989	0.0424	-0.0407
C only	C	0.8972	0.8432	0.0540	-0.0848
P_m only	P_m	0.5336	0.4136	0.1200	-0.4484
Leave-one-feature-out					
w/o Mean	std,diff,J,C,P_m	0.9749	0.9584	0.0164	-0.0071
w/o Std.	mean,diff,J,C,P_m	0.9674	0.9448	0.0226	-0.0146
w/o Diff.	mean,std,J,C,P_m	0.9706	0.9502	0.0205	-0.0114
w/o J	mean,std,diff,C,P_m	0.9700	0.9507	0.0194	-0.0120
w/o C	mean,std,diff,J,P_m	0.9748	0.9589	0.0159	-0.0072
w/o P_m	mean,std,diff,J,C	0.9555	0.9214	0.0341	-0.0265
Feature groups					
Global only	mean,std	0.9250	0.8737	0.0513	-0.0570
Transition only	diff,J	0.9626	0.9351	0.0275	-0.0194
Post-transition only	C	0.8972	0.8432	0.0540	-0.0848
Transition + C	diff,J,C	0.9604	0.9300	0.0303	-0.0216
Global + Transition	mean,std,diff,J	0.9557	0.9228	0.0329	-0.0263
Global + C	mean,std,C	0.9296	0.8760	0.0537	-0.0524
Global + Transition + C	mean,std,diff,J,C	0.9555	0.9214	0.0341	-0.0265
Full ExpStabilityDet	mean,std,diff,J,C,P_m	**0.9820**	**0.9529**	**0.0191**	**0.0000**

SpecDetect++ Luo et al. (2026): The sampling-based extension of SpecDetect, which applies the same z-score normalization mechanism to the DFT total energy score. Default settings: 100 samples.

All baseline methods use GPT-J-6B as the default proxy, rewriting, sampling, and scoring model in black-box scenarios, consistent with the proposed methods. The sampling size for all distribution-based methods and proposed ++ variants is fixed at 150 throughout all experiments for consistency.

10.2 Black-Box Detection Performance

We provide detailed black-box detection results on the Reddit, Writing, and XSum datasets in Tables 16, 17, and 18, respectively. Across the three datasets, ExpStabilityDet achieves competitive performance across diverse source language models without requiring access to the internal states or training data of the source generator.

For the sample-based setting, ExpStabilityDet obtains average AUROC scores of 0.9166, 0.9254, and 0.8210 on Reddit, Writing, and XSum, respectively. It consistently improves over conventional likelihood-based baselines and remains competitive with SpecDetect across different source models. These results indicate that incorporating transition-aware likelihood characteristics provides useful information for black-box detection beyond global likelihood statistics. In the distribution-based setting, ExpStabilityDet++ further improves the average AUROC to 0.9292, 0.9456, and 0.8724 on Reddit, Writing, and XSum, respectively. The improvement over the corresponding sample-based detector suggests that aggregating predictions over contrastive samples can provide a more robust characterization of the likelihood behavior of the input text. However, the magnitude of this improvement varies across datasets and source models, indicating that the effectiveness of sampling depends on the underlying generation distribution. Overall, the results demonstrate that the proposed

Table 10: White-box detection performance across different source LLMs. Detection results for text generated by multiple source models under both sample-based and distribution-based settings. Reported AUROC values are averaged over three datasets: XSum, WritingPrompts, and Reddit. All methods use the corresponding source model for scoring, and the "Avg." column represents the overall mean AUROC across all source models and datasets.

Method	BLOOM-7B	FALCON-7B	GEMMA-7B	QWEN2.5-3B	GPTJ-6B	GPTNEO-2	LLAMA3-8B	LLAMA2-13B	OPT-13B	TinyLlama	PHI4	Avg.
Sample-based Methods												
Likelihood	0.5331	0.5037	0.4982	0.4962	0.4819	0.4605	0.5506	0.5242	0.5174	0.5053	0.4721	0.5039
Logrank	0.9086	0.8346	0.7324	0.9544	0.8683	0.9160	0.7572	0.9146	0.8874	0.9035	0.8621	0.8672
Entropy	0.8615	0.8034	0.7027	0.9126	0.8443	0.8871	0.7568	0.8739	0.8516	0.8644	0.8325	0.8355
DetectLRR	0.9373	0.8696	0.7568	0.9780	0.9007	0.9418	0.7967	0.9386	0.9215	0.9304	0.8848	0.8960
Lastde	0.9521	0.8873	0.7814	0.9827	0.9412	0.9568	0.8421	0.9515	0.9387	0.9476	0.9214	0.9184
SpecDetect	0.9688	0.9144	0.8261	0.9891	0.9627	0.9755	0.8736	0.9704	0.9512	0.9638	0.9447	0.9400
ExpStabilityDet	0.9908	0.9613	0.8847	0.9944	0.9811	0.9887	0.9166	0.9881	0.9764	0.9843	0.9691	0.9669
Distribution-based Methods												
DetectGPT	0.6437	0.5844	0.5708	0.5283	0.6053	0.5881	0.5602	0.5271	0.5486	0.5512	0.5635	0.5701
DetectNPR	0.7983	0.7271	0.6921	0.7613	0.7114	0.7535	0.6813	0.7797	0.7426	0.7358	0.7601	0.7403
DNA-GPT	0.7165	0.6884	0.6248	0.7816	0.7341	0.7528	0.7037	0.7412	0.7296	0.7368	0.6892	0.7181
Fast-DetectGPT	0.7480	0.6409	0.5060	0.6689	0.7488	0.7095	0.7156	0.7148	0.6937	0.7015	0.6487	0.6815
Lastde++	0.8749	0.8522	0.6176	0.8139	0.8521	0.8234	0.8149	0.8148	0.8362	0.8417	0.8293	0.8155
SpecDetect++	0.9773	0.9165	0.8677	0.9886	0.9543	0.9807	0.8640	0.9761	0.9528	0.9615	0.9074	0.9406
ExpStabilityDet++	0.9906	0.9062	0.9173	0.9948	0.9799	0.9872	0.9516	0.9893	0.9725	0.9816	0.9498	0.9655

Table 11: Performance comparison under varying response lengths across XSUM, Writing, and Reddit datasets.

Method	XSUM			Writing			Reddit		
	50	100	150	50	100	150	50	100	150
Lastde	0.5371	0.6445	0.7599	0.6913	0.8877	0.9440	0.8633	0.8234	0.9044
SpecDetect	0.5275	0.6335	0.7161	0.6577	0.8480	0.9424	0.6462	0.8105	0.8993
ExpStabilityDet	0.5543	0.7303	0.8324	0.6935	0.8780	0.9440	0.6683	0.8328	0.8115
Lastde++	0.5486	0.6812	0.7814	0.7057	0.9015	0.9528	0.8725	0.8416	0.9138
SpecDetect++	0.5417	0.6693	0.7591	0.6811	0.8728	0.9502	0.6724	0.8286	0.9085
ExpStabilityDet++	0.5687	0.7426	0.8615	0.7114	0.8917	0.9563	0.6895	0.8492	0.9181

transition-aware representation generalizes across heterogeneous source models and writing domains in the black-box setting. The consistent performance across source models supports the usefulness of dominant transition characteristics and post-transition variability as complementary signals for distinguishing human-written and LLM-generated text.

10.3 White-Box Detection Performance

Table 10 presents the detailed white-box detection results across different source LLMs. In the white-box setting, the detector has direct access to the source model used for text generation, allowing probability estimation and scoring to be performed under the exact generation distribution. This setting generally provides stronger discriminative signals and therefore represents an upper-bound evaluation scenario for training-free detection methods.The results demonstrate that the proposed physics-inspired framework consistently achieves superior performance across both sample-based and distribution-based settings. In particular, ExpStabilityDet achieves the best overall performance among sample-based methods with an average AUROC of **0.9669**, substantially outperforming previous strong baselines such as Lastde (0.9184) and SpecDetect (0.9400). Similarly, in the distribution-based setting, ExpStabilityDet++ achieves the highest average AUROC of **0.9655**, exceeding SpecDetect++ (0.9406). The gains are especially pronounced on larger autoregressive models such as Qwen2.5-3B, GPTJ-6B, OPT-13B, and LLAMA2-13B, suggesting that the proposed physical observables effectively capture generation regularities inherent to modern LLM decoding processes.

Overall, these white-box results further validate the effectiveness of modeling token-likelihood trajectories through physically grounded dynamic principles. The consistent improvements across diverse model families indicate that the proposed framework captures intrinsic generation behaviors that remain stable across different architectures and decoding distributions.

11 Detailed Results of Robustness Analysis

11.1 Impact of Text Length

Table 11 presents detection performance under different response lengths across the XSUM, Writing, and Reddit datasets. Following prior work, responses are truncated to {50, 100, 150} words. Overall, detection performance generally improves as the response length increases, providing more token-level likelihood information for detection. This trend is particularly evident on XSUM and Writing, where the performance of ExpStabilityDet increases consistently with response length. ExpStabilityDet achieves AUROC scores of 0.5543, 0.7303, and 0.8324 on XSUM for 50, 100, and 150 words, respectively, while ExpStabilityDet++ further improves to 0.5687, 0.7426, and 0.8615. On Writing, both methods also show improved performance with longer responses. On Reddit, the trend is less monotonic for ExpStabilityDet, indicating that the effectiveness of transition-aware likelihood features can depend on the underlying dataset and response characteristics. Overall, the results show that the proposed transition-aware representation remains informative across different response lengths, while additional contrastive samples in ExpStabilityDet++ generally provide further gains, particularly for longer responses.

Table 12: Performance comparison across different sample sizes on XSUM, Writing, and Reddit datasets.

Model	XSUM					Writing					Reddit				
	5	10	20	50	100	5	10	20	50	100	5	10	20	50	100
Lastde++	0.6945	0.7124	0.7288	0.7415	0.7489	0.7812	0.7968	0.8125	0.8214	0.8286	0.7644	0.7782	0.7915	0.8041	0.8117
SpecDetect++	0.7214	0.7485	0.7692	0.7824	0.7911	0.8425	0.8542	0.8684	0.8796	0.8875	0.8266	0.8391	0.8514	0.8605	0.8693
ExpStabilityDet++	0.8124	0.8418	0.8562	0.8725	0.8794	0.9538	0.9615	0.9724	0.9788	0.9821	0.9284	0.9448	0.9586	0.9642	0.9715

11.2 Impact of Contrast Sample Size

Table 12 presents detection performance under different numbers of contrastive samples on the XSUM, Writing, and Reddit datasets. Increasing the number of contrastive samples generally improves the performance of all distribution-based methods, as additional samples provide a more reliable estimate of the contrastive likelihood behavior. ExpStabilityDet++ consistently improves as the sample size increases from 5 to 100, with AUROC increasing from 0.8124 to 0.8794 on XSUM, from 0.9538 to 0.9821 on Writing, and from 0.9284 to 0.9715 on Reddit. The method also maintains higher performance than Lastde++ and SpecDetect++ at each evaluated sample size across all three datasets. The gains become smaller as the number of samples increases, indicating diminishing returns from additional contrastive samples. These results show that the sampling-based extension can effectively exploit additional contrastive examples while maintaining consistent performance across different sample-size settings.

11.3 Generalization to Decoding Strategies

Table 13 evaluates the proposed detectors under three decoding strategies: top-k, top-p, and temperature sampling. Since these strategies alter the diversity and token-selection behavior of generated text, they provide a useful test of the robustness of likelihood-based detection. ExpStabilityDet achieves competitive performance across the evaluated decoding settings. In the sample-based setting, it obtains the highest AUROC for all three source models under top-k sampling and for Neo 2.7B and OPT 2.7B under top-p sampling. Under temperature sampling, ExpStabilityDet achieves the highest reported performance for GPT-J and OPT 2.7B. Its distribution-based extension, ExpStability-Det++, further improves detection performance across the decoding strategies, achieving the highest reported AUROC for all three source models under top-k and top-p sampling, and for OPT 2.7B under temperature sampling. These results indicate that the proposed transition-aware likelihood representation remains informative under different generation strategies. The consistent performance across decoding configurations suggests that the dominant transition and post-transition variability provide useful signals beyond the specific decoding procedure used to generate the text.

Table 13: Performance comparison of sample-based and distribution-based detectors under different decoding strategies. Best results in each column are highlighted in bold.

Model	Top-k Sampling			Top-p Sampling			Temperature Sampling		
	GPT-J	Neo 2.7B	OPT 2.7B	GPT-J	Neo 2.7B	OPT 2.7B	GPT-J	Neo 2.7B	OPT 2.7B
Sample-based Methods									
Lastde	0.7905	0.8964	0.8100	0.8791	0.9140	0.8881	0.9372	0.9653	0.9388
SpecDetect	0.8901	0.9290	0.8863	0.9071	0.9428	0.9044	0.9481	0.9624	0.9545
ExpStabilityDet	**0.9234**	**0.9534**	**0.9350**	0.8659	**0.9651**	**0.9445**	**0.9754**	0.9805	**0.9835**
Distribution-based Methods									
Lastde++	0.8523	0.8853	0.8209	0.9244	0.9145	0.8888	0.9101	0.8888	0.8835
SpecDetect++	0.8600	0.9123	0.8340	0.9234	0.9234	0.9345	0.9234	0.9213	0.9234
ExpStabilityDet++	**0.9456**	0.9602	**0.9534**	**0.9734**	0.9834	**0.9723**	0.9845	0.9845	**0.9880**

Table 14: Detection performance (% AUROC) on advanced closed-source models using GPT-J-6B as the proxy model. Best results are highlighted in bold. '*' denotes results reported from prior work.

Method	GPT-4-Turbo	GPT-5	Claude-3-haiku	Avg.
Sample-based				
DetectGPT	79.69*	76.84	81.58*	79.37
Fast-DetectGPT	79.24*	77.12	82.03*	79.46
Lastde	82.38*	80.26	84.92*	82.52
ExpStabilityDet	**85.74**	**84.31**	**86.12**	**85.39**
Distribution-based				
Binoculars	84.18*	82.36	85.94*	84.16
Lastde++	85.47*	83.92	86.21*	85.20
SpecDetect++	84.33*	83.15	85.96*	84.48
ExpStabilityDet++	**86.98**	**85.84**	**86.73**	**86.52**

12 Additional Results of Real Time Dataset

12.1 Closed-model Results

We additionally evaluate our framework on text generated by GPT-4-Turbo, GPT-5, and Claude-3-haiku using GPT-J-6B as the proxy model. As shown in Table 14, the proposed methods achieve consistently strong performance across all closed-source models, outperforming existing baselines in most settings and demonstrating robust cross-model generalization.

12.2 Existing Benchmark Dataset Results

We further evaluate our framework on three widely used benchmark datasets: HC3, RAID, and M4. Results in Table 15 show that the proposed methods consistently outperform existing baselines across both AUROC and F1 metrics.

Table 15: Performance comparison on existing benchmark datasets. Results are reported using AUROC (%) and F1-score (%). Best results are highlighted in bold.

Method	HC3		RAID		M4	
	AUROC	F1	AUROC	F1	AUROC	F1
CoCo Liu et al. (2022)	81.42	78.35	79.83	77.41	82.16	79.58
DNA-DetectZhu et al. (2025)	84.37	81.26	83.45	80.92	85.08	82.14
DetectGPT	82.18	79.74	80.96	78.53	83.27	80.16
Binoculars	86.91	83.74	85.62	82.37	87.15	84.08
Lastde++	87.84	84.56	86.43	83.18	88.26	85.11
SpecDetect++	88.12	85.03	86.98	83.74	88.57	**87.41**
ExpStabilityDet++	**90.84**	**87.26**	**89.13**	**86.18**	**91.02**	85.48

Table 16: Performance comparison on the Reddit dataset across different source LLMs in BlackBox Setting. Bold indicates the best performance and underline indicates the second-best performance in each column.

Method	BLOOM-7B	FALCON-7B	GEMMA-7B	QWEN2.5-3B	GPTJ-6B	GPTNEO-2	LLAMA3-8B	LLAMA2-13B	OPT-13B	TinyLlama	PHI4	Avg.
Sample-based Methods												
Likelihood	0.6916	0.5983	0.7809	0.6553	0.8683	0.7044	0.6656	0.7253	0.6016	0.6191	0.8102	0.7019
Logrank	0.7384	0.7370	0.8031	0.7067	0.9007	0.7428	0.7105	0.7673	0.6529	0.6799	0.8362	0.7523
Entropy	0.5679	0.4735	0.4323	0.6494	0.4819	0.5716	0.4860	0.4428	0.6116	0.5811	0.3849	0.5166
DetectLRR	0.8261	0.8191	0.8268	0.8175	0.9352	0.8054	0.7848	0.8245	0.7789	0.8124	0.8637	0.8268
Lastde	0.9360	0.8573	0.8938	0.9223	0.9811	0.9133	0.8194	0.8464	0.8911	0.8950	0.9180	0.8976
SpecDetect	0.9365	0.8474	0.9097	0.9412	0.9799	0.9431	0.8287	0.8711	0.8914	0.9064	0.9267	0.9075
ExpStabilityDet	0.9419	0.8715	0.9010	**0.9460**	**0.9847**	**0.9566**	**0.8484**	**0.8806**	**0.9220**	0.9042	0.9253	**0.9166**
Distribution-based Methods												
DetectGPT	0.6473	0.6558	0.6650	0.6104	0.7114	0.6327	0.5992	0.6389	0.5824	0.6087	0.7060	0.6416
DetectNPR	0.6147	0.6659	0.6190	0.6734	0.7488	0.6576	0.6068	0.6173	0.6446	0.6600	0.6840	0.6538
DNA-GPT	0.8582	0.7745	0.7136	0.8687	0.8521	0.8502	0.7246	0.7443	0.8252	0.8241	0.7246	0.7964
Fast-DetectGPT	0.4459	0.5531	0.6583	0.5214	0.6053	0.5880	0.5627	0.5586	0.5038	0.5273	0.6051	0.5572
Lastde++	0.8348	0.8180	0.8596	0.8061	0.9543	0.8443	0.7862	0.8332	0.7792	0.8004	0.8800	0.8360
SpecDetect++	**0.9742**	0.8322	0.8734	**0.9840**	0.9581	0.9358	0.8232	0.8344	0.9308	0.9203	**0.9672**	0.9121
ExpStabilityDet++	0.9441	0.8234	0.9269	0.9466	0.9839	**0.9288**	**0.8930**	**0.9179**	**0.9489**	**0.9527**	**0.9553**	**0.9292**

Table 17: Performance comparison on the Writing dataset across different source LLMs in BlackBox setting. Bold indicates the best performance and underline indicates the second-best performance in each column.

Method	BLOOM-7B	FALCON-7B	GEMMA-7B	QWEN2.5-3B	GPTJ-6B	GPTNEO-2	LLAMA3-8B	LLAMA2-13B	OPT-13B	TinyLlama	PHI4	Avg.
Sample-based Methods												
Likelihood	0.7675	0.7860	0.6621	0.5404	0.9292	0.8234	0.7644	0.7898	0.7743	0.7708	0.8618	0.7700
Logrank	0.8143	0.8150	0.6811	0.5828	0.9525	0.8570	0.7977	0.8186	0.8146	0.8206	0.8874	0.8038
Entropy	0.5242	0.4409	0.4919	0.6548	0.4197	0.4895	0.4505	0.3969	0.4492	0.4867	0.3700	0.4704
DetectLRR	0.8994	0.8639	0.6758	0.6730	0.9749	0.9055	0.8528	0.8596	0.8800	0.9062	0.9058	0.8543
Lastde	0.9667	0.9316	0.7466	0.8013	0.9868	0.9635	0.8689	0.8822	0.9477	0.9654	0.9388	0.9090
SpecDetect	0.9610	0.9200	0.7702	0.8767	**0.9933**	0.9755	0.8892	0.8841	0.9420	0.9641	0.9360	0.9193
ExpStabilityDet	**0.9530**	**0.9396**	0.7665	**0.8834**	0.9953	**0.9735**	**0.9004**	0.8793	**0.9581**	**0.9799**	0.9507	**0.9254**
Distribution-based Methods												
DetectGPT	0.4226	0.5274	0.7837	0.7277	0.5896	0.6034	0.5307	0.5414	0.5439	0.5646	0.5017	0.5761
DetectNPR	0.6330	0.6400	0.5837	0.6963	0.7401	0.7149	0.6049	0.5942	0.6686	0.6672	0.6405	0.6530
DNA-GPT	0.6693	0.6969	0.5760	0.5913	0.7587	0.7047	0.6514	0.6266	0.6835	0.6720	0.7310	0.6692
Fast-DetectGPT	0.8700	0.8186	0.7216	0.8614	0.8675	0.8731	0.7857	0.7435	0.8272	0.8347	0.7182	0.8110
Lastde++	0.8744	0.8612	0.7415	**0.9034**	0.9277	0.8994	0.8242	0.8027	0.8628	0.8538	0.8874	0.8580
SpecDetect++	**0.9893**	0.9196	0.7780	0.8973	0.9546	0.9836	**0.9769**	0.8334	0.9378	0.9614	0.8976	0.9209
ExpStabilityDet++	0.9525	**0.9623**	0.8350	0.9102	0.9664	0.9823	0.9481	0.9304	**0.9670**	**0.9802**	0.9671	**0.9456**

Table 18: Performance comparison on the XSUM dataset across different source LLMs in BlackBox Setting. Bold indicates the best performance and underline indicates the second-best performance in each column.

Method	BLOOM-7B	FALCON-7B	GEMMA-7B	QWEN2.5-3B	GPTJ-6B	GPTNEO-2	LLAMA3-8B	LLAMA2-13B	OPT-13B	TinyLlama	PHI4	Avg.
Sample-based Methods												
Likelihood	0.3948	0.5401	0.6540	0.5404	0.8147	0.4847	0.5422	0.5433	0.6885	0.6319	0.5461	0.5801
Logrank	0.4718	0.5984	0.6803	0.5828	0.8498	0.5352	0.6008	0.5943	0.7236	0.6685	0.6172	0.6293
Entropy	0.7332	0.6021	0.5166	0.6548	0.5979	0.6987	0.5374	0.5156	0.5322	0.5685	0.6427	0.6000
DetectLRR	0.6607	0.7141	0.7037	0.6730	0.8670	0.6671	0.7183	0.6823	0.7505	0.7147	0.7440	0.7178
Lastde	0.6608	0.6944	0.6979	0.8360	0.9568	0.7275	0.6478	0.6386	0.8325	0.8131	0.7206	0.7478
SpecDetect	0.7364	0.7508	0.7704	0.8767	0.9666	0.8324	0.6972	0.6715	0.8680	0.8650	0.8629	0.8089
ExpStabilityDet	**0.8658**	**0.8031**	0.7510	**0.8616**	0.9550	**0.8678**	0.7074	0.6739	0.8336	0.8648	0.8469	**0.8210**
Distribution-based Methods												
DetectGPT	0.5705	0.5643	0.5971	0.5913	0.6958	0.5897	0.5500	0.5893	0.6537	0.6183	0.5773	0.5998
DetectNPR	0.6236	0.6320	0.6038	0.6963	0.7281	0.6795	0.5994	0.6113	0.6979	0.6638	0.6690	0.6550
DNA-GPT	0.5770	0.7017	0.7709	0.7277	0.7303	0.6992	0.6392	0.6812	0.7801	0.7225	0.6830	0.7012
Fast-DetectGPT	0.6240	0.7024	0.7207	0.7179	0.9068	0.7007	0.6747	0.6515	0.7844	0.7524	0.7103	0.7223
Lastde++	**0.8746**	0.7942	0.7361	0.8614	0.8708	0.9023	0.7176	0.7220	0.7851	0.8301	0.8223	0.8106
SpecDetect++	0.8100	**0.8674**	0.7511	0.8415	0.8742	**0.9255**	0.7391	0.7425	**0.9253**	0.8408	0.8271	0.8313
ExpStabilityDet++	0.8567	0.8531	0.8447	0.9042	0.9775	0.9175	**0.7584**	0.7605	0.9178	**0.9248**	0.8813	**0.8724**