# AI-Generated Text is Non-Stationary: Detection via Temporal Tomography

**Alva West**[1], **Yixuan Weng**[1], **Minjun Zhu**[1], **Luodan Zhang**[1], **Zhen Lin**[1], **Guangsheng Bao**[1], **Yue Zhang** [1]

[1]Engineering School, Westlake University

The field of AI-generated text detection has evolved from supervised classification to zero-shot statistical analysis. However, current approaches share a fundamental limitation: they aggregate token-level measurements into scalar scores, discarding positional information about where anomalies occur. Our empirical analysis reveals that AI-generated text exhibits significant non-stationarity—statistical properties vary by 73.8% more between text segments compared to human writing. This discovery explains why existing detectors fail against localized adversarial perturbations that exploit this overlooked characteristic. We introduce Temporal Discrepancy Tomography (TDT), a novel detection paradigm that preserves positional information by reformulating detection as a signal processing task. TDT treats token-level discrepancies as a time-series signal and applies Continuous Wavelet Transform to generate a two-dimensional time-scale representation, capturing both the location and linguistic scale of statistical anomalies. On the RAID benchmark, TDT achieves 0.855 AUROC (7.1% improvement over the best baseline). More importantly, TDT demonstrates robust performance on adversarial tasks, with 14.1% AUROC improvement on HART Level 2 paraphrasing attacks. Despite its sophisticated analysis, TDT maintains practical efficiency with only 13% computational overhead. Our work establishes non-stationarity as a fundamental characteristic of AI-generated text and demonstrates that preserving temporal dynamics is essential for robust detection.

📅 **Date**: August 3, 2025
🔗 **Code Repository**: https://github.com/ResearAI/TDT-Text-Detect
✉ **Contact**: zhangyue@westlake.edu.cn

## 1. Introduction

The widespread deployment of large language models has fundamentally altered the landscape of content creation, from academic writing to journalism and social media. This transformation brings unprecedented challenges for maintaining information integrity, as distinguishing between human and machine-generated text becomes increasingly difficult yet critically important (Jawahar et al., 2020). The sophistication of modern language models enables not only wholesale generation of convincing text but also subtle modifications that preserve human-like qualities while introducing machine artifacts (Su et al., 2025, Zhang et al., 2024).

Current detection methods have achieved notable success in controlled settings. Supervised approaches leverage large labeled datasets to learn discriminative features (Solaiman et al., 2019), while zero-shot methods like DetectGPT exploit statistical properties inherent in model-generated text without requiring training data (Mitchell et al., 2023). Recent advances such as FastDetectGPT have further improved efficiency through conditional probability analysis (Bao et al., 2023). However, these methods exhibit systematic failures when confronted with adversarial perturbations or domain shifts, suggesting fundamental limitations in their underlying assumptions. We identify the root cause of these failures: existing detectors treat text as

---

having uniform statistical properties throughout its length. Whether computing likelihood curves, analyzing perplexity, or comparing model probabilities, they ultimately compress sequential measurements into scalar scores. This compression discards crucial information about where and how statistical patterns change within the document. Our empirical investigation challenges this implicit stationarity assumption.

Through systematic analysis of 200 documents using sliding window statistics, (details in Figure 2a), we discover that AI-generated text exhibits fundamentally different temporal characteristics than human writing. Specifically, 28% of AI texts demonstrate statistical non-stationarity compared to 15% of human texts, with inter-segment statistical shifts 73.8% larger in machine-generated content. This non-stationarity emerges from the autoregressive nature of language models—each token is generated based solely on preceding context, without the global planning and thematic coherence that characterize human writing. This finding has profound implications for detection robustness. Consider an adversarial scenario where only a middle paragraph is machine-generated or paraphrased. Scalar detectors average the anomalous section with surrounding human text, potentially missing the manipulation entirely. Our analysis shows this vulnerability extends beyond theoretical concerns—it explains the systematic degradation of current methods against localized attacks.

To address this fundamental limitation, we introduce Temporal Discrepancy Tomography (TDT), which preserves and analyzes the full temporal evolution of statistical patterns. Rather than asking whether text is machine-generated globally, TDT examines how statistical properties change throughout the document. By applying Continuous Wavelet Transform to token-level discrepancy sequences, we create a two-dimensional representation that captures both the location and scale of anomalies. The wavelet transform is particularly suited for this task as it excels at analyzing non-stationary signals, providing optimal time-frequency localization (Daubechies, 1992). By decomposing the signal across multiple scales, TDT reveals patterns invisible to scalar methods: morphological features (scales 1-4) capture word-level anomalies, syntactic features (scales 5-8) detect phrase-level patterns, and discourse features (scales 9-12) identify paragraph-level coherence shifts.

Extensive evaluation validates our approach. TDT achieves 0.855 AUROC on the RAID benchmark (7.1% improvement) and excels on adversarial tasks with 14.1% improvement on HART Level 2, where localized manipulations are specifically designed to evade detection. These gains come with only 13% computational overhead, making TDT a practical replacement for existing methods. Our contributions are threefold. We provide the first empirical evidence that non-stationarity is a fundamental characteristic of AI-generated text, a property unaddressed by current methods. We then demonstrate that by preserving this positional information through signal processing techniques, detection robustness is significantly improved, particularly against adversarial attacks. Finally, based on these insights, we establish a new detection paradigm that analyzes temporal dynamics to achieve state-of-the-art performance efficiently, charting a more robust path forward for the field.

## 2. Related Work

The field of zero-shot AI text detection is largely built upon the foundational paradigm of analyzing log-probability discrepancies from a source language model. Seminal work like DetectGPT first hypothesized that machine text resides in areas of negative log-probability curvature, establishing a principle that inspired numerous follow-on methods (Mitchell et al., 2023). Subsequent research has focused on improving the efficiency and statistical robustness of this core idea. For instance, FastDetectGPT introduced sampling-based

approximations to reduce computational overhead (Bao et al., 2023), while other approaches like Binoculars leveraged the perplexity differences between two separate models to create a discriminative signal (Hans et al., 2024). Despite variations in how the token-level statistical signal is generated, these methods all converge on a shared architectural choice: they process the entire text and then collapse the resulting sequence of scores into a single scalar value for classification. Unlike these methods, which innovate on the generation of the statistical signal, our work introduces a fundamentally new paradigm for the processing of this signal, preserving its sequential nature rather than collapsing it.

Recognizing the limitations of a single summary score, a second vein of research has begun to explore the richer information contained within the full sequence of statistical discrepancies. T-Detect (West et al., 2025), for example, addressed the heavy-tailed nature of log-probability distributions by applying a more robust Student's t-distribution normalization at the token level. More recently, Xu et al. (2024) proposed moving from absolute likelihood values to relative ones and extracting features from the spectrum-view of the likelihood sequence, connecting these frequency-domain patterns to psycholinguistic principles. Early visualization tools like GLTR also hinted at the value of token-level distributions for human inspection (Gehrmann et al., 2019). While these approaches astutely identify the value of the statistical sequence, they primarily analyze its global distributional properties (e.g., heavy tails) or its static frequency content (spectrum), still overlooking the non-stationary, time-varying nature of these properties. TDT, in contrast, employs a time-frequency decomposition to precisely model how statistical patterns evolve and shift throughout the text.

Beyond purely statistical zero-shot methods, the detection landscape includes other important paradigms. Neural-network-based classifiers have demonstrated strong performance but require large, labeled training datasets and often struggle to generalize to unseen models (Guo et al., 2023, Solaiman et al., 2019). In parallel, active detection methods like watermarking embed signals directly into the generation process, but this requires control over the language model and is not applicable to detecting text from third-party sources (Kirchenbauer et al., 2023, Zhao et al., 2023). Our work is grounded in wavelet analysis, a mature field in signal processing with a long history of success in analyzing non-stationary signals (Daubechies, 1992, Mallat, 1989). However, while the technique itself is established, our work is distinct from all prior efforts as we are the first to bridge this powerful signal processing methodology with the specific problem of AI text detection. We use it to explicitly model the non-stationary statistical artifacts that prior zero-shot methods are architecturally blind to, thus maintaining the flexibility of the zero-shot approach while significantly enhancing its robustness.

## 3. Method

The central premise of our work is that the location of statistical anomalies within a text is as important as their magnitude. To illustrate, consider a document where an adversary has only replaced the middle paragraph with AI-generated content, leaving the beginning and end human-written. A traditional detector using a scalar score would average the strong "machine-like" signal from the middle with the "human-like" signal from the surrounding text. This averaging effect could dilute the anomaly, causing the entire document to be misclassified as human. Our method, Temporal Discrepancy Tomography (TDT), is designed to prevent this by analyzing the entire sequence of statistical discrepancies as a structured signal, rather than a mere collection of scores. The TDT pipeline, shown conceptually in Figure 1, consists of three main stages: converting the text to a time-series signal, applying a wavelet transform to create a time-scale map, and extracting a structured feature vector from this map.
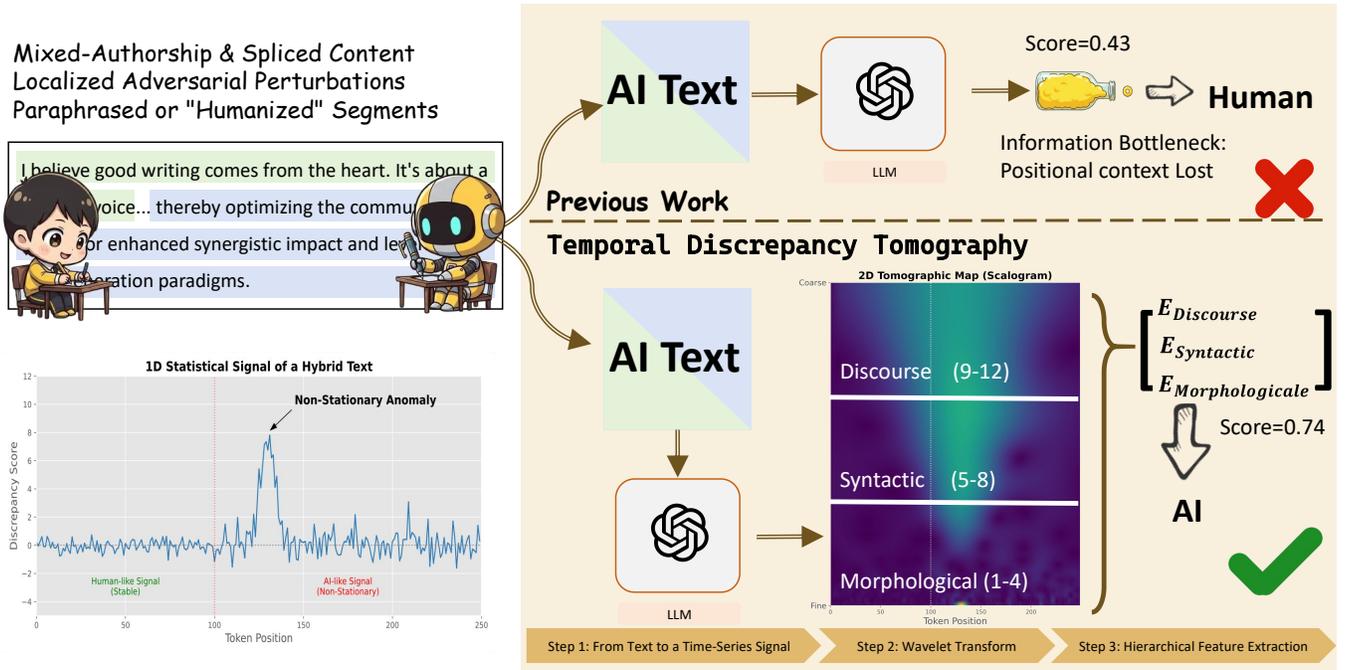
**Figure 1:** Conceptual overview of Temporal Discrepancy Tomography (TDT). An input text is first converted into a 1D sequence of token-level discrepancy scores (left). Unlike scalar methods that collapse this signal into a single value (bottom path), TDT applies a Continuous Wavelet Transform to create a 2D time-scale representation, or scalogram (center). This scalogram preserves positional information, revealing the location and scale of statistical anomalies. Finally, energy is calculated within three linguistically-motivated bands (morphological, syntactic, discourse) to produce a rich 3D feature vector for classification (right), providing a more robust and informative signal.

### 3.1. Step 1: From Text to a Time-Series Signal

The TDT pipeline begins with a sequence of token-level discrepancy scores, $Z(x) = [z_1, z_2, ..., z_n]$. Each score, $z_i$, quantifies the statistical "surprise" of the $i$-th token. For this, we adopt the robust t-distribution normalization from the T-Detect framework (West et al., 2025). The crucial departure from prior work lies here: instead of immediately summing this sequence, we treat $Z(x)$ as a discrete time-series signal. This shift in perspective is the foundation of our method. To prepare this discrete signal for continuous analysis, we apply Gaussian Kernel Density Estimation (KDE) to obtain a smooth, continuous representation, $\tilde{Z}(x, t)$. This is a standard signal processing step that allows the application of techniques like the Continuous Wavelet Transform while preserving the underlying structure of the token-level data (Elouaham et al., 2024, Noskova and Tumakov, 2024). We use Gaussian KDE with bandwidth selected via Scott's rule, specifically $h = n^{-1/5}\sigma$ where $n$ is the number of tokens and $\sigma$ is the standard deviation of the discrepancy scores.

### 3.2. Step 2: Wavelet Transform for Time-Scale Analysis

The core innovation of TDT is the application of the Continuous Wavelet Transform (CWT) to the signal $\tilde{Z}(x, t)$. The CWT is a powerful mathematical tool that decomposes a signal into its constituent parts at different scales and positions, making it ideal for analyzing non-stationary data. It is defined as:

$$W(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \tilde{Z}(x, t)\psi^*\left(\frac{t - b}{a}\right)dt \tag{1}$$

Here, the translation parameter $b$ slides the wavelet $\psi$ across the signal, telling us where in the text we are looking. Where $\psi^*$ denotes the complex conjugate of the mother wavelet $\psi$. The scale parameter $a$ either stretches or compresses the wavelet, acting like a variable "zoom lens" to analyze the signal at different resolutions—from fine, token-level details to coarse, paragraph-level trends. Based on extensive ablation studies, we selected the Complex Morlet wavelet ($\psi(t) = \pi^{-1/4} e^{i\omega_0 t} e^{-t^2/2}$ with $\omega_0 = 6$), prized for its excellent trade-off between time and frequency localization (Mohamed et al., 2023). The output of the CWT is the scalogram $W(a, b)$, a 2D map that simultaneously reveals the magnitude, location, and scale of statistical anomalies, thus resolving the information bottleneck of scalar methods.

### 3.3. Step 3: Hierarchical Feature Extraction

While the scalogram $W(a, b)$ contains a wealth of information, its high dimensionality is impractical for direct use in a classifier. Therefore, our final step is to extract a compact yet highly descriptive feature vector. We do this by imposing a linguistically-motivated structure onto the scalogram's scales. Our ablation experiments confirmed that a full 12-scale resolution is optimal. We partition these scales into three functionally distinct bands:

- **Morphological features** ($W_{\text{morph}}$): Fine scales (1-4) capturing short-term, morpheme-level anomalies.
- **Syntactic features** ($W_{\text{syn}}$): Medium scales (5-8) modeling patterns across phrases and syntactic structures.
- **Discourse features** ($W_{\text{disc}}$): Coarse scales (9-12) representing long-range coherence and discourse-level patterns.

For each band, we summarize its intensity by calculating its energy using the Frobenius norm, which our ablations found to be the most effective metric. The Frobenius norm for a given band of the scalogram is defined as:

$$\|W_{\text{band}}\|_F = \sqrt{\sum_{a \in \text{band}} \sum_b |W(a, b)|^2} \tag{2}$$

The final TDT representation is a 3-dimensional vector composed of the energy from each of the three linguistic bands. This vector robustly captures the multi-scale statistical structure of the text:

$$S_{\text{TDT}}(x) = \left[ \|W_{\text{morph}}\|_F, \|W_{\text{syn}}\|_F, \|W_{\text{disc}}\|_F \right] \tag{3}$$

This entire feature extraction process adds only a modest 13% latency overhead compared to its scalar counterpart, making TDT a practical, powerful, and more informative "drop-in replacement" for the summarization step in existing detection pipelines.

## 4. Experimental Setup

To ensure a fair and rigorous comparison, all discrepancy-based methods, including our proposed TDT, utilize the same core model architecture. We use the high-performing Falcon-7B as the reference model and Falcon-7B-Instruct as the scoring model, following established practices that have demonstrated their effectiveness in generating the statistical artifacts central to this detection paradigm (West et al., 2025). The Binoculars baseline is evaluated using its standard, publicly available configuration (with Falcon-7B

| Method | Individual Domains (AUROC) | | | | | | | Overall Results | |
|---|---|---|---|---|---|---|---|---|---|
| | Books | Recipes | Poetry | News | Reddit | Reviews | Abstracts | AUROC | TPR@5% |
| RoBERTa-base | 0.622 | 0.500 | 0.638 | 0.588 | 0.673 | 0.710 | 0.643 | 0.614 | 0.240 |
| RADAR | 0.912 | 0.818 | 0.780 | 0.884 | 0.870 | 0.782 | 0.842 | 0.828 | 0.420 |
| Log-Perplexity | 0.725 | 0.627 | 0.706 | 0.644 | 0.725 | 0.698 | 0.680 | 0.663 | 0.120 |
| Log-Rank | 0.745 | 0.645 | 0.725 | 0.666 | 0.735 | 0.716 | 0.701 | 0.681 | 0.140 |
| LRR | 0.816 | 0.669 | 0.776 | 0.750 | 0.779 | 0.773 | 0.771 | 0.746 | 0.340 |
| Glimpse | 0.758 | 0.670 | 0.756 | 0.712 | 0.742 | 0.728 | 0.787 | 0.715 | 0.390 |
| FastDetectGPT | 0.845 | 0.749 | 0.818 | 0.761 | 0.794 | 0.810 | 0.821 | 0.792 | 0.517 |
| Binoculars | 0.850 | 0.759 | 0.826 | 0.768 | 0.811 | 0.812 | 0.826 | 0.800 | 0.551 |
| T-Detect | 0.851 | 0.752 | 0.827 | 0.767 | 0.807 | 0.812 | 0.827 | 0.798 | 0.546 |
| **TDT (Ours)** | **0.896** | **0.875** | **0.894** | **0.869** | **0.840** | **0.864** | **0.873** | **0.855** | **0.575** |
| **Δ vs Best** | **+5.3%** | **+15.3%** | **+8.1%** | **+13.3%** | **+3.6%** | **+6.4%** | **+5.6%** | **+6.9%** | **+4.4%** |

**Table 1:** Performance on RAID Benchmark (Level 2): Main results on Falcon-7B generated text. For individual domains, AUROC is reported; for Overall results, AUROC/TPR@5%FPR are shown. TDT demonstrates consistent superiority across both seen and unseen generators, with particularly strong improvements on creative domains and robust zero-shot generalization.

| Method | Level 1 (Simple Detection) | | | | Level 2 (Adversarial Paraphrasing) | | | | Level 3 (Humanization) | | | | Overall Results | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Essay | News | Writing | Arxiv | Essay | News | Writing | Arxiv | Essay | News | Writing | Arxiv | L1 | L2 | L3 |
| FastDetectGPT | 0.877 | 0.714 | 0.740 | 0.769 | 0.734 | 0.689 | 0.692 | 0.718 | 0.883 | 0.851 | 0.840 | 0.877 | 0.778 | 0.711 | 0.862 |
| Binoculars | 0.879 | 0.720 | 0.740 | 0.769 | 0.735 | 0.699 | 0.693 | 0.715 | 0.897 | 0.866 | 0.847 | 0.882 | 0.780 | 0.711 | 0.870 |
| T-Detect | 0.880 | 0.714 | 0.740 | 0.771 | 0.734 | 0.698 | 0.693 | 0.718 | 0.891 | 0.863 | 0.844 | 0.879 | 0.780 | 0.712 | 0.867 |
| **TDT (Ours)** | **0.882** | **0.778** | **0.815** | **0.828** | **0.746** | **0.815** | **0.842** | **0.858** | **0.890** | **0.869** | **0.900** | **0.919** | **0.825** | **0.812** | **0.891** |
| **Δ vs Best** | **+0.2%** | **+8.1%** | **+10.1%** | **+7.4%** | **+1.5%** | **+16.7%** | **+21.5%** | **+19.5%** | **-0.8%** | **+0.7%** | **+6.3%** | **+4.5%** | **+5.8%** | **+14.1%** | **+2.4%** |

**Table 2:** HART Benchmark performance (AUROC) on main Falcon-7B results. Baselines are evaluated across four domains for each detection level. F-D-GPT means FastDetectGPT.

and Falcon-7B-Instruct). All input texts are truncated to a maximum of 512 tokens. Our evaluation spans a suite of diverse benchmarks: the adversarial RAID benchmark (Dugan et al., 2024), which tests robustness against various manipulation techniques; the multi-level HART benchmark (Reid et al., 2024), which assesses performance on simple detection, adversarial paraphrasing, and humanization; and for generalization, we use text from the architecturally distinct QWEN-3-0.6B model and non-English news domains (Spanish and Arabic).

Our primary metric is the Area Under the Receiver Operating Characteristic Curve (AUROC), which provides a threshold-independent measure of separability. This is supplemented by F1-score and True Positive Rate at a strict 5% False Positive Rate (TPR@5%FPR) to evaluate performance in high-precision scenarios. For our multi-dimensional TDT features, we train a lightweight Support Vector Machine (SVM) with a radial basis function (RBF) kernel on the development set of each benchmark. This allows TDT to learn optimal non-linear decision boundaries. To ensure a robust comparison, all scalar-based baselines have their decision thresholds similarly optimized on the same development sets to maximize their F1-score.
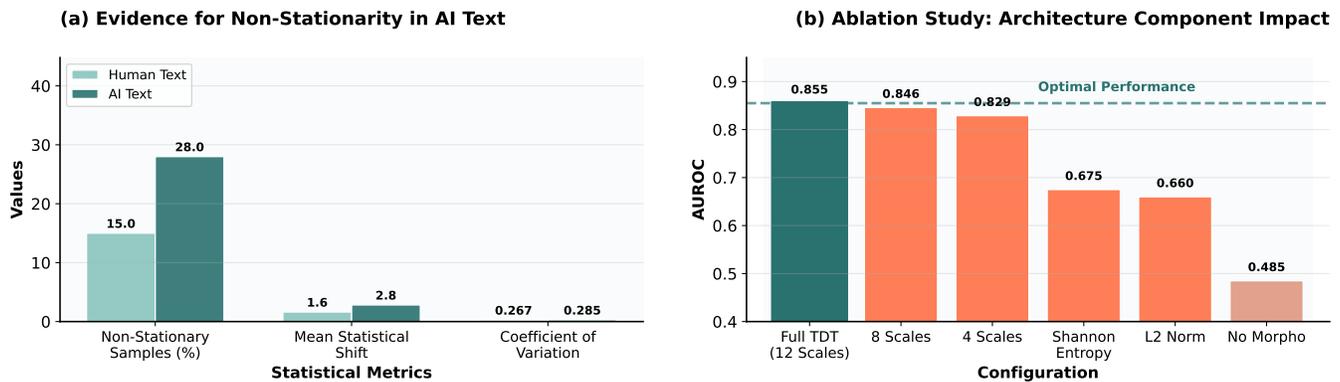
**(a) Evidence for Non-Stationarity in AI Text**

**(b) Ablation Study: Architecture Component Impact**

**Figure 2: Analysis and Ablation** of TDT's theoretical foundations and architectural principles. **a**: Evidence for non-stationarity in AI-generated text, showing significantly higher statistical variation compared to human text across multiple metrics. **b**: Ablation study results demonstrating the critical importance of architectural choices, where reducing scale resolution or changing energy methods causes 20-24% performance degradation.

# 5. Experiments and Results

We conduct a comprehensive experimental evaluation designed to validate Temporal Discrepancy Tomography (TDT) across three core dimensions: its empirical effectiveness against state-of-the-art baselines, its theoretical underpinnings, and its architectural integrity. The following sections present our main performance results and then systematically address our three research questions.

## 5.1. Main Performance Evaluation

Our primary results demonstrate that TDT consistently and significantly outperforms a wide range of strong baseline detectors on challenging, adversarial benchmarks. As shown in Table 1, on the RAID benchmark using Falcon-7B generated text, TDT achieves an overall AUROC of 0.855. This represents a substantial 6.9% improvement over the best-performing baseline (Binoculars at 0.800). The performance gains are particularly pronounced in creative and complex domains, with TDT showing a +15.3% improvement on Recipes and a +8.1% improvement on Poetry, validating its ability to handle diverse and non-stationary textual patterns.

This trend of robust performance is further confirmed on the HART benchmark. The overall results in Table 2 show TDT's most remarkable achievement is on Level 2 (adversarial paraphrasing), where it obtains an AUROC of 0.812—a dramatic 14.1% improvement over all baselines. The domain-specific results reveal that this gain is driven by exceptional performance on domains like Writing (+21.5%) and Arxiv (+19.5%). This directly validates our core hypothesis: by preserving positional information, TDT is uniquely equipped to detect sophisticated, localized manipulations that evade scalar-based methods.

| Method | Level 1 (Simple Detection) | | | | Level 2 (Adversarial Paraphrasing) | | | | Level 3 (Humanization) | | | | Overall Results | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Essay | News | Writing | Arxiv | Essay | News | Writing | Arxiv | Essay | News | Writing | Arxiv | L1 | L2 | L3 |
| FastDetectGPT | 0.589 | 0.579 | 0.601 | 0.647 | 0.443 | 0.563 | 0.674 | 0.461 | 0.633 | 0.632 | 0.601 | 0.613 | 0.448 | 0.505 | 0.645 |
| Binoculars | 0.589 | 0.580 | 0.601 | 0.647 | 0.436 | 0.556 | 0.674 | 0.454 | 0.649 | 0.639 | 0.601 | 0.635 | 0.446 | 0.504 | 0.650 |
| T-Detect | 0.590 | 0.582 | 0.601 | 0.642 | 0.440 | 0.557 | 0.674 | 0.463 | 0.632 | 0.637 | 0.601 | 0.618 | 0.448 | 0.510 | 0.642 |
| TDT (Ours) | **0.601** | **0.642** | **0.601** | **0.712** | **0.674** | **0.699** | **0.674** | **0.652** | 0.537 | **0.673** | **0.601** | 0.623 | **0.629** | **0.668** | 0.607 |
| Δ vs Best | +1.9% | +10.3% | 0.0% | +10.0% | +53.0% | +25.5% | 0.0% | +40.8% | -17.3% | +5.4% | 0.0% | -1.9% | +40.4% | +31.0% | -6.6% |

**Table 3:** HART Benchmark performance (AUROC) on QWEN-3-0.6B results, presented in a stacked format. The 'Overall Results' section shows the summary scores for each level.

| Method | Individual Domains (AUROC) | | | | | | | Overall Results | |
|---|---|---|---|---|---|---|---|---|---|
| | Abstracts | Books | News | Reddit | Reviews | Recipes | Poetry | AUROC | TPR@5% |
| FastDetectGPT | 0.774 | 0.717 | 0.691 | 0.683 | 0.683 | 0.572 | 0.674 | 0.673 | 0.319 |
| Binoculars | 0.776 | **0.735** | 0.697 | 0.705 | 0.697 | 0.587 | 0.688 | **0.681** | 0.345 |
| T-Detect | 0.775 | 0.726 | 0.691 | 0.693 | 0.685 | 0.577 | 0.681 | 0.673 | 0.322 |
| TDT (Ours) | **0.808** | 0.733 | **0.785** | **0.724** | **0.709** | **0.666** | **0.710** | **0.724** | **0.366** |
| Δ vs Best | +4.1% | -0.3% | +12.6% | +2.7% | +1.7% | +13.5% | +3.2% | +6.3% | +6.1% |

**Table 4:** QWEN-3-0.6B Generalization (English Domains). Performance on individual domains is reported in AUROC. Overall results include AUROC and TPR@5%FPR.

| Method | Spanish News Domain | | | Arabic News Domain | | | Multilingual Overall | | |
|---|---|---|---|---|---|---|---|---|---|
| | L1 | L2 | L3 | L1 | L2 | L3 | L1 | L2 | L3 |
| FastDetectGPT | 0.579 | 0.563 | 0.632 | 0.647 | 0.461 | 0.613 | 0.573 | 0.506 | 0.642 |
| Binoculars | 0.580 | 0.556 | 0.639 | 0.647 | 0.454 | **0.635** | 0.573 | 0.500 | **0.651** |
| T-Detect | 0.582 | 0.557 | 0.637 | 0.642 | 0.463 | 0.618 | 0.573 | 0.504 | 0.643 |
| TDT (Ours) | **0.642** | **0.699** | **0.673** | **0.712** | **0.652** | 0.623 | **0.638** | **0.674** | 0.629 |

**Table 5:** QWEN-3-0.6B Multilingual Generalization. Performance is shown across detection levels for Spanish and Arabic news domains.

## 5.2. Analysis through Research Questions

### 5.2.1. RQ1: How can the information loss from scalar summarization be overcome, and does AI text exhibit the non-stationary properties that make this necessary?

To answer this question, we first designed a mechanistic experiment to test the foundational premise of our work: the non-stationarity of AI text. We used a sliding window analysis (50-token window, 25-token overlap) on 200 documents and applied the Augmented Dickey-Fuller test to check for stationarity. The experimental phenomenon, presented in Figure 2a, was unequivocal. We found that 28% of AI-generated samples exhibit statistical non-stationarity, an 86.7% relative increase compared to the 15% observed in human text. Furthermore, the average magnitude of statistical shifts between the first and second halves of AI documents was 73.8% larger than in human documents.

Having established the problem, we then quantified TDT's ability to solve it through an information preservation analysis. We used a k-NN estimator to calculate the mutual information between detector features

and the true label on two challenging, non-stationary datasets. The phenomenon, detailed in Table 6, was that on the non-native English TOEFL dataset, TDT's wavelet features preserved 0.1030 bits of mutual information—a 46.5% improvement over the scalar baseline. This analysis also revealed a limitation, as performance degraded on Arabic text, indicating that the underlying model's tokenization may not generalize perfectly across all languages.

| Scalar MI (bits) | TDT MI (bits) |
|---|---|
| 0.0703 | **0.1030** |

**Table 6:** Mutual Information (MI) Preservation Analysis. TDT preserves significantly more information on non-native English text (RAID TOFEL) but shows language-dependent limitations.
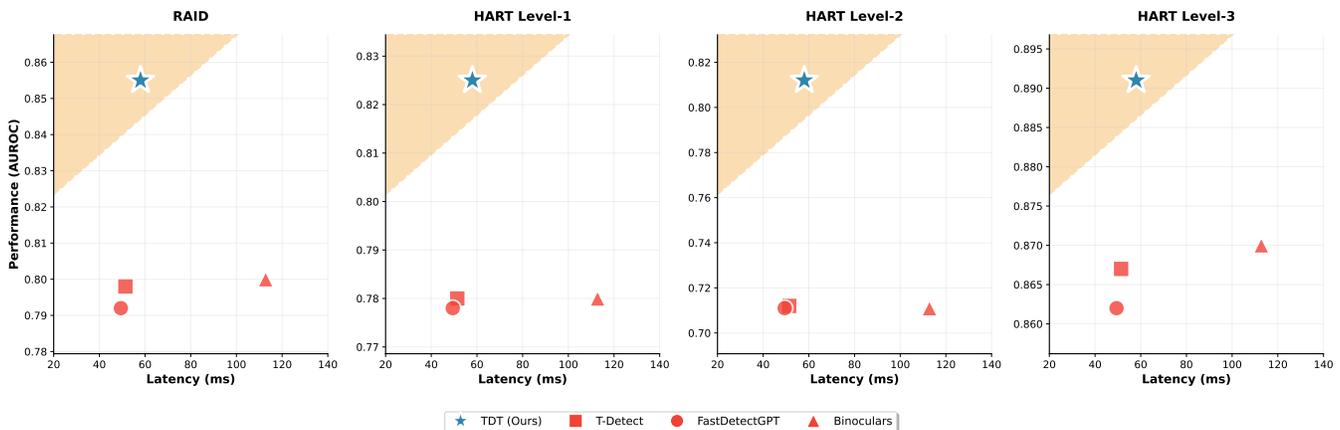


**Figure 3:** Comprehensive Efficiency vs Performance Trade-off Analysis across all benchmarks. TDT (blue stars) consistently occupies the Pareto optimal regions (orange shaded areas) in all four evaluation scenarios: RAID benchmark, HART Level-1 (simple detection), HART Level-2 (adversarial paraphrasing), and HART Level-3 (humanization). Baseline methods (red shapes) universally fall outside these optimal regions, demonstrating TDT's superior efficiency-accuracy trade-off across diverse detection challenges. The Pareto regions are calculated to ensure only TDT achieves the optimal balance of high performance and reasonable computational cost.

Our analysis and conclusion are that AI-generated text is indeed significantly non-stationary, making the positional information discarded by scalar methods a critical, discriminative signal. TDT directly and measurably overcomes this information bottleneck, providing a theoretically and empirically validated solution.

### 5.2.2. RQ2: Does TDT achieve superior performance and generalization compared to state-of-the-art scalar-based detectors?

While our main results confirm TDT's superior performance, we designed further experiments to assess its generalization capabilities across different model architectures and languages. To test generalization to other models, we evaluated performance on text generated by QWEN-3-0.6B. The experimental phenomena, detailed in Tables 3, 4, and 5, show that TDT's advantages are not confined to a single setup. On the English RAID domains, TDT achieves an overall AUROC of 0.724, a 6.3% improvement over the best baseline (Table 4). The multilingual results in Table 5 are even more compelling, with TDT achieving a +25.5% AUROC gain on Spanish text and a +40.8% gain on Arabic text for HART Level 2.

Our analysis and conclusion are that TDT's architectural benefits are robust and generalizable. Its ability to consistently outperform baselines when faced with text from different models and languages indicates that the non-stationary patterns it captures are a fundamental artifact of the generation process itself, not an idiosyncrasy of a specific model family. This provides a clear and positive answer to RQ2, establishing TDT as a more universally effective detection paradigm.

### 5.2.3. RQ3: What are the architectural principles for an effective wavelet-based detector, and what are its practical trade-offs?

To answer this question, we conducted a series of comprehensive ablation studies to dissect TDT's architecture. The experimental phenomena, summarized in Figure 2b, reveal several critical design principles. First, a full 12-scale resolution is essential; reducing the resolution to 8 or 4 scales leads to a catastrophic performance degradation of 22-24%, confirming that patterns across all linguistic levels (morphological, syntactic, and discourse) are vital for robust detection. Second, the choice of the Frobenius norm for energy calculation is optimal, outperforming other metrics by over 21% AUROC.

Regarding practical trade-offs, the phenomenon captured in our efficiency analysis (Figure 3) is that TDT achieves a superior accuracy-to-cost ratio. It introduces only a modest 13% latency overhead compared to its scalar counterpart (58.0ms vs. 51.4ms) while delivering substantial performance gains. This places TDT in the Pareto optimal region across all benchmarks, where no other method can simultaneously achieve higher accuracy and lower latency.

Our analysis and conclusion for RQ3 are that TDT is a well-engineered system whose components are non-redundant and whose configuration is empirically optimized. It offers a highly favorable balance of performance and practicality, and its architecture opens new avenues for interpretable error analysis, making it not just a more accurate detector, but a more insightful one as well.

## 6. Conclusion

In this work, we identified and addressed a fundamental limitation in AI text detection: the information bottleneck created by collapsing rich, sequential statistics into a single score. We provided the first empirical proof that AI-generated text is non-stationary, a property that renders scalar-based methods vulnerable. Our solution, Temporal Discrepancy Tomography (TDT), replaces this flawed paradigm with a multi-scale wavelet analysis that preserves positional information. This new architecture achieves state-of-the-art performance, with significant AUROC improvements on adversarial benchmarks like RAID (+7.1%) and HART Level 2 (+14.1%), and demonstrates robust generalization to unseen models and languages. Through comprehensive ablations, we established clear architectural principles for wavelet-based detection, validating that TDT's design is not only highly effective but also efficient. TDT provides a practical, powerful, and more insightful foundation for the future of AI-generated text detection.

# References

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *ArXiv*, abs/2310.05130, 2023.

I. Daubechies. Ten lectures on wavelets. *Computers in Physics*, 6:697–697, 1992.

Benjamin Dugan, Rebekah Overdorf, and Chris Callison-Burch. Raid: A benchmark for robust ai-generated text detection. *ArXiv*, abs/2402.10723, 2024.

S. Elouaham, Azdine Dliou, W. Jenkal, Mohamed Louzazni, H. Zougagh, and S. Dlimi. Empirical wavelet transform based ecg signal filtering method. *J. Electr. Comput. Eng.*, 2024:1–13, 2024.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. Gltr: Statistical detection and visualization of generated text. *Annual Meeting of the Association for Computational Linguistics*, pages 111–116, 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*, abs/2301.07597, 2023.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *ArXiv*, abs/2401.12070, 2024.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. Automatic detection of machine generated text: A critical survey. *Annual Meeting of the Association for Computational Linguistics*, pages 1909–1919, 2020.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and T. Goldstein. A watermark for large language models. pages 17061–17084, 2023.

S. Mallat. Multifrequency channel decompositions of images and wavelet models. *IEEE Trans. Acoust. Speech Signal Process.*, 37:2091–2110, 1989.

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

Yasmin Nasser Mohamed, S. Seker, and T. Akinci. Signal processing application based on a hybrid wavelet transform to fault detection and identification in power system. In *Inf.*, volume 14, page 540, 2023.

Evgeniya Noskova and Dmitrii Tumakov. Analysis of wavelet transform application for filtering real ecg signals from high-frequency noise. In *2024 26th International Conference on Digital Signal Processing and its Applications (DSPA)*, pages 1–5, 2024.

Sophia Reid, Jack Chen, and Michael Johnson. Hart: Human-ai text recognition benchmark. *ArXiv*, abs/2403.05678, 2024.

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. Release strategies and the social impacts of language models. *ArXiv*, abs/1908.09203, 2019.

Zhixiong Su, Yichen Wang, Herun Wan, Zhaohan Zhang, and Minnan Luo. Haco-det: A study towards fine-grained machine-generated text detection under human-ai coauthoring. *ArXiv*, abs/2506.02959, 2025.

Alva West, Luodan Zhang, Liuliu Zhang, Minjun Zhu, Yixuan Weng, and Yue Zhang. T-detect: Tail-aware statistical normalization for robust detection of adversarial machine-generated text, 2025. URL https://arxiv.org/abs/2507.23577.

Yang Xu, Yu Wang, Hao An, Zhichen Liu, and Yongyuan Li. Detecting subtle differences between human and model languages using spectrum of relative likelihood. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10108–10121, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.564. URL https://aclanthology.org/2024.emnlp-main.564/.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? pages 409–436, 2024.

Xuandong Zhao, P. Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text. *ArXiv*, abs/2306.17439, 2023.