TriBet: Relative E-values for Online Machine-Generated Text Detection

Xunye Tian[1,2*] **Zhijian Zhou**[1,*] **Liuhua Peng**[1] **Jared Collette**[2]
Dino Sejdinovic[3] **Yue Yang**[2] **Feng Liu**[1,†]

[1]The University of Melbourne [2]Maincode [3]Adelaide University
fengliu.ml@gmail.com

Abstract

To ensure a formal guarantee that fully human-written text streams are *not falsely accused* as machine-generated text (MGT), the online betting framework (e-values) has been successfully applied on various existing detectors. However, current two-sample e-value methods fail in two real-world application scenarios: the false positive rate (FPR) is often out of control and the test power drops drastically when (1) the reference domain is similar but not exactly same as test domain (cross-domain), or (2) proxy model of detector is outdated compared to the modern source model. We identify the root cause as an online calibration dilemma: two-sample optimization and betting relies on a null-calibrating offset that is unidentifiable from unlabeled test streams under shift. We resolve this with a triple-sample betting framework (TriBet) that conduct online MGT detection as relative testing against both human and machine references using a witness score, and introduces a conformal test-inclusive domain-shift betting strategy to achieve FPR control without fragile online calibration. Under round-wise conditionally exchangeable calibration, TriBet is anytime-valid at level α; asymptotic power one and stopping-time bounds hold under additional detectability and log-growth conditions. Finite-pool implementations are evaluated empirically. Empirically, we evaluate on 15 source LLMs, 8 proxy models and mixed-document streams, TriBet consistently delivers faster and more powerful detection than both oracle and online betting baselines while keeping FPR strictly controlled.

1 Introduction

Large language models (LLMs) can now produce fluent text across collaborative story writing [1], educational content [2], and news summarisation [3]. This progress also leads to a negative side of misuse of LLMs, such as neural fake news [4], misinformation [5], plagiarism [6], academic cheating [7, 8] and the broader societal harms of foundation-model deployment [9]. To help safeguard the trustworthiness of information systems, a substantial body of work has emerged on machine-generated text (MGT) detection [10]. Existing MGT detection methods broadly include zero-shot statistical detectors, supervised classifiers and watermarking methods. Since watermarking embeds a signal at the generation time and requires cooperation from the source model [11], it is outside the post-hoc black-box audit setting considered in this work. We therefore focus on post-hoc detectors: **statistical zero-shot detectors** that compute likelihood- or perturbation-based signals from a proxy LLM [12–15], and **supervised binary classifier detectors** fine-tuned on labelled human-written text (HWT) and MGT corpora [16, 17]. State-of-the-art instances of each family report 90–99 % per-text accuracy at low false-positive rate (FPR) on early-era short-text benchmarks (100–500 tokens) [13, 15, 17].

*Equal contribution. †Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

As source LLMs continue to scale, however, the distributional gap between machine-generated and human-written text on standard divergence measures has been observed to shrink [18, 19], and a line of work has shown that, in the limiting regime where this gap vanishes, the AUROC of any per-text classifier is upper-bounded by a quantity approaching the random-guessing value of 0.5 [18]; complementary results characterise the sample-complexity regime in which detection remains feasible [20]. Per-text decisions on individual short spans are therefore increasingly unreliable on their own, motivating a shift from isolated fixed-sample detection to document-level evidence accumulation.

Online MGT detection. In real deployment, however, MGT-bearing content rarely arrives as a closed, fixed-size batch. News websites, social platforms, and long commercial reports publish content sequentially, and a single document already contains many spans: a typical paragraph spans roughly 100 tokens, so a single page contains 5–10 spans and a 200-page commercial report contains hundreds. An offline detector that is correct on human spans with probability a (i.e. per-span FPR $q = 1 - a$) then yields, for a document of n spans, a document-level false-accusation rate of $1 - a^n$ that approaches certainty for moderate n (Figure 1): even at the optimistic per-span accuracy $a = 0.95$, a 20-span document is falsely accused with probability $1 - 0.95^{20} \approx 0.64$, and at $n = 200$ the probability exceeds 0.99996. To address this scaling problem, Chen and Wang [21] introduce *online* MGT detection via sequential testing-by-betting [22–24]: at each round t they maintain a non-negative wealth process W_t that is a supermartingale under H_0 (the document is purely HWT), built from a two-sample payoff function f and an online Newton step (ONS) strategy [25] for betting on $f(y_t) - \varepsilon$ with a null-calibrating offset ε. Ville's inequality then bounds $\Pr(\exists t : W_t \geq 1/\alpha \mid H_0) \leq \alpha$, meaning that the probability of ever crossing the rejection threshold under a purely human stream is at most α, uniformly over stopping times. The document-level FPR is therefore held at the target level α regardless of stream length, and the test rejects at the first t for which $W_t \geq 1/\alpha$. Closely related kernel-MMD and Huber e-process variants [26, 27] fall in the same template. The evidence-accumulating online detection setting is therefore more reasonable at deployment scale: it is what *converts strong per-span detection signal into a document-level decision with the formal FPR guarantee* needed by downstream audit systems.

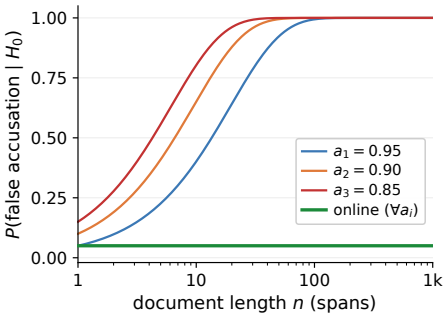


Figure 1: Document-level false-accusation rate against document length of n spans. Coloured curves: three offline detectors with HWT accuracy a_1, a_2, a_3, each compounding to $1 - a_i^n$. Green: using *any* of the three detectors into an online betting framework yields the same anytime-α line at the target level α, by Ville's inequality, whenever the assumptions hold; the framework is detector-agnostic at the FPR level.

Method motivation. Existing online MGT detection typically calibrate a scalar detector score through an offset ε, thereby coupling power and false-positive control. In the construction of Chen and Wang [21], this offset can only be estimated from the offline human references, but it is difficult to precisely identify in two scenarios that are commonly tested in real life. First, under *cross-domain* scenario, the pre-owned human reference samples and the audited stream may come from similar but distinct domains, such as using a general dataset XSum [28] as a reference while the testing stream is from the Olympic news [29]; any within-reference estimate of ε is blind to the testing-stream domain shift. Second, under *weak detection scores*, the alternative payoff can fall below the offline estimate of ϵ, causing the wealth process to stagnate even when the stream contains MGT.

Empirically, the online variant of Chen and Wang [21] violates anytime-α control on 7 of 9 detector–source-model settings under DetectGPT [12] / NPR[14], with empirical FPR up to 0.81 on cross-domain online scenario. We refer to this obstruction as the *online-calibration dilemma*: within the ONS-with-ε betting strategy, the offset that determines the cross-domain power is also the quantity that must be calibrated for false-positive control. This creates a fragile dependence on estimating ε from the offline human reference under the domain shift or score misspecification. We resolve this obstacle by decoupling evidence construction from anytime calibration.

Our solution. We propose TRIBET, a triple-sample online betting framework for post-hoc MGT detection. The key design principle is to decouple two roles that are entangled in prior online betting detectors: constructing evidence that is powerful against MGT, and calibrating this evidence to guarantee anytime level-α FPR control. Inspired from relative hypothesis testing [30, 31], TRIBET

compares each incoming text jointly against human and machine references through a *triple-sample relative kernel witness function*, yielding a bounded score that indicates whether the text is closer to the human or machine reference distribution. To calibrate this evidence, TRIBET replaces offset estimation with a *test-inclusive domain-shift conformal e-process*, so validity follows from conformal calibration and permutation invariance rather than assumptions on the raw detector score distribution.

Contribution. Our framework design makes TRIBET compatible with a broad class of offline detectors while directly addressing the cross-domain and score-misspecification regimes that break existing online betting methods. In abundant experiments, we show TRIBET fully solved the dilemma faced by baselines, and exhibits robust power gains and strict FPR control over online betting baselines across diverse significance level, LLMs, proxy detectors, and domains.

2 Preliminaries

Problem setup. We study the online MGT detection problem. At each round $t \geq 1$, a query text y_t arrives sequentially from an unknown source P_Y. The task is to detect the presence of machine-generated text as early as possible, while ensuring that a fully human-written stream is falsely flagged with probability at most a pre-specified level α. We test the following hypotheses:

$$H_0 : \text{ the online stream is purely HWT}, \qquad H_1 : \text{ the online stream contains MGT}. \tag{1}$$

A valid online detector should reject H_0 quickly when H_1 is true, while controlling the probability of ever rejecting H_0 under a purely human-written stream by α.

Sequential hypothesis testing by betting. Sequential testing by betting [22, 23] tests H_0 by constructing a nonnegative supermartingale $(W_t)_{t\geq 0}$ under H_0: $\mathbb{E}[W_t \mid \mathcal{F}_{t-1}] \leq W_{t-1}$, where $\mathcal{F}_{t-1}$ is the information available before round t. By Ville's inequality,

$$\Pr\big(\exists t \geq 1 : W_t \geq 1/\alpha \mid H_0\big) \leq \alpha. \tag{2}$$

Therefore, stopping and rejecting at the first t when $W_t \geq 1/\alpha$ gives a level-α sequential testing.

Sequential two-sample mean testing for MGT detection. Chen and Wang [21] formalize the MGT detection problem in (1) as a sequential two-sample mean testing problem. They first map each text y to a scalar detector score $\phi(y)$, under which HWT and MGT are expected to be distinguishable. Given a human-written reference distribution P_X, the resulting score-level hypotheses are

$$H_0^{2\text{s}} : |\mathbb{E}_{x\sim P_X}[\phi(x)] - \mathbb{E}_{y\sim P_Y}[\phi(y)]| \leq \varepsilon, \qquad H_1^{2\text{s}} : |\mathbb{E}_{x\sim P_X}[\phi(x)] - \mathbb{E}_{y\sim P_Y}[\phi(y)]| > \varepsilon\,,$$

where $\varepsilon \in \mathbb{R}$ is a *null-calibration offset*. Given this formulation, the online stream is tested through a scalar payoff constructed from the detector score. At round t, given a human reference sample $x_t \sim P_X$, define payoffs $f(y_t) \in \{\phi(x_t) - \phi(y_t), \phi(y_t) - \phi(x_t)\}$ [2] , where larger values indicate stronger evidence toward MGT in the corresponding score direction.

The wealth process is then updated by betting on the centered payoff $f(y_t) - \varepsilon$:

$$W_t = W_{t-1}\big(1 + \lambda_t\big(f(y_t) - \varepsilon\big)\big)\,, \quad W_0 = 1, \tag{3}$$

where λ_t is a predictable *betting fraction* with $\lambda_t \in [0, 1/(2d_t)]$ and $d_t \geq |f(y_t) - \varepsilon|$. The d_t-bound keeps W_t positive, while λ_t is updated by the Online Newton Step (ONS) algorithm [24, 25, 32] to maximize wealth growth.

This construction has two key components. The first is the *payoff* f, which determines the evidence strength under $H_1^{2\text{s}}$. The second is the *calibration offset* ε. To make the score-level testing problem yield a valid test for the original null H_0 in (1), the offset must satisfy the null-validity condition

$$\mathbb{E}^{H_0^{2\text{s}}}_{y\sim P_Y}[f(y)] \leq \varepsilon.$$

Here, for simplicity, we write $\mathbb{E}^{H_0^{2\text{s}}}_{y\sim P_Y}[f(y)]$ for $|\mathbb{E}_{x\sim P_X}[\phi(x)] - \mathbb{E}_{y\sim P_Y}[\phi(y)]|$ under $H_0^{2\text{s}}$, and similarly define $\mathbb{E}^{H_1^{2\text{s}}}_{y\sim P_Y}[f(y)]$ under $H_1^{2\text{s}}$. In Section 3, we show that the null-validity condition is fragile under domain shift and weakly discriminative score. In Section 4, we propose TriBet, which formalizes MGT detection in (1) as a relative testing problem by comparing each query against both human and machine references. It replaces the two-sample payoff and offset-based betting with a triple-sample kernel-witness payoff and a conformal e-process, without relying on an offset ε.

[2] The hypotheses use an absolute mean discrepancy, so both payoffs are tested by two e-processes at levels $\alpha/2$ and $\alpha/2$.

Table 1: Representative failures of classical two-sample ONS-calibrated betting. Columns report the valid-band quantities and mean power/FPR over 100 repetitions at $\alpha = 0.05$. The upper block shows below-band FPR inflation; the lower block shows power collapse from above-band or empty-band behavior. Markers denote FPR $> \alpha$ (†) and near-zero power (‡).

Failure mode	Source LLM	Detector (Proxy)	$\mathbb{E}^{H_0^{2s}}_{y\sim P_Y}[f(y)]$	$\hat{\varepsilon}$	$\mathbb{E}^{H_1^{2s}}_{y\sim P_Y}[f(y)]$	(Pow, FPR)
FPR inflation (below band)	PaLM 2	DetectGPT (Gemma-2B)	0.336	0.214	0.840	(1.00, **0.40**†)
	Gemini-1.5-Flash	DetectGPT (Gemma-2B)	0.354	0.248	0.715	(1.00, **0.59**†)
	Gemini-1.5-Pro	NPR (Gemma-2B)	0.025	0.014	0.048	(1.00, **0.81**†)
Power collapse (above band)	GPT-5.4	FDGPT (Gemma3-12B)	0.396	0.844	0.423	(**0.02**‡, 0.02)
	Claude-Sonnet-4.6	FDGPT (Falcon-7B)	0.574	0.564	0.492	(**0.02**‡, 0.04)

3 The online-calibration dilemma

The wealth process in (3) imposes two competing requirements on the calibration offset ε. For false-positive control, i.e., to ensure that a purely human-written stream is falsely rejected with probability at most α, the wealth process must remain a supermartingale under H_0 as in (2), which requires $\mathbb{E}^{H_0^{2s}}_{y\sim P_Y}[f(y)] \leq \varepsilon$. To achieve test power, i.e., to correctly reject H_0 under H_1 with high probability, the wealth process should grow under the alternative H_1, which requires $\mathbb{E}^{H_1^{2s}}_{y\sim P_Y}[f(y)] > \varepsilon$. Thus, useful values of ε must lie in the valid band

$$\mathbb{E}^{H_0^{2s}}_{y\sim P_Y}[f(y)] \leq \varepsilon < \mathbb{E}^{H_1^{2s}}_{y\sim P_Y}[f(y)]\,. \tag{4}$$

How source-only estimation $\hat{\varepsilon}$ can leave the band. In the online MGT detection scenario, Chen and Wang [21] estimate $\hat{\varepsilon}$ using an offline human reference set, before observing the target testing stream. This source-only estimation reflects the payoff behavior in the reference domain, but it has no direct access to the human distribution of the online stream. Therefore, when the testing stream differs from the reference domain, the estimated offset may not fall inside the valid band in (4).

First, under *cross-domain shift*, the condition for false-positive control depends on the testing-stream human distribution, which may differ from the offline reference domain used to estimate $\hat{\varepsilon}$. Since $\hat{\varepsilon}$ is estimated only from the source reference set, it can be too small, i.e., $\hat{\varepsilon} < \mathbb{E}^{H_0^{2s}}_{y\sim P_Y}[f(y)]$, thereby violating the supermartingale condition and inflating false positives. Second, under *heavy-tailed or weakly discriminative scores*, the estimate can be unstable and the valid band can become narrow. In this case, $\hat{\varepsilon}$ can be relative large, i.e., $\mathbb{E}^{H_1^{2s}}_{y\sim P_Y}[f(y)] \leq \hat{\varepsilon}$, so the wealth process no longer grows reliably under H_1 and the test loses power. Together, these two cases show why a single source-only offset is fragile under domain shift or weakly discriminative scores: the offset is estimated from the source reference domain, whereas the valid band is determined by the testing-stream human null and the MGT alternative. We refer to this mismatch as the *online-calibration dilemma*.

Empirical findings. Table 1 illustrates the two failure modes predicted by the valid-band diagnosis on representative source-LLM, proxy-detector, and score-function combinations. We report the source-only estimate $\hat{\varepsilon}$, the testing-stream null mean $\mathbb{E}^{H_0^{2s}}_{y\sim P_Y}[f(y)]$, and the alternative mean $\mathbb{E}^{H_1^{2s}}_{y\sim P_Y}[f(y)]$, so that the position of $\hat{\varepsilon}$ relative to the valid band can be checked directly. The upper block shows false-positive failures: $\hat{\varepsilon}$ falls below the null mean, i.e., $\mathbb{E}^{H_0^{2s}}_{y\sim P_Y}[f(y)] > \hat{\varepsilon}$, so the supermartingale condition is violated and the empirical FPR exceeds the nominal level $\alpha = 0.05$. The lower block shows power-collapse failures: $\hat{\varepsilon}$ lies above the alternative mean, i.e., $\mathbb{E}^{H_1^{2s}}_{y\sim P_Y}[f(y)] \leq \hat{\varepsilon}$, or the valid band is nearly empty, so the wealth process cannot grow reliably under H_1. These failures motivate the next section: instead of estimating a single source-only offset, we construct relative evidence using both human and machine references and calibrate it without estimating ε.

4 TriBet: A triple-sample betting framework

TriBet aims to provide anytime level-α validity without relying on the calibration offset ε. It formalizes MGT detection in (1) as a relative testing problem by comparing each query against both human and machine references. Specifically, let $\mathcal{X} = \{x_i\}_{i=1}^{m}$ be a human-written reference set

drawn from P_X, and let $\mathcal{Z} = \{z_j\}_{j=1}^n$ be a machine reference set generated offline by the same proxy LLM used by the detector score ϕ. For any text distribution P, we write $P_{\phi(U)}$ for the distribution of the detector score $\phi(U)$ when $U \sim P$.

Let ρ be a discrepancy measure between distributions on $\mathbb{R}$. We formulate MGT detection in (1) as a relative testing problem in the detector-score space:

$$H_0^{\text{rel}} : \rho(P_{\phi(Y)}, P_{\phi(X)}) \leq \rho(P_{\phi(Y)}, P_{\phi(Z)}), \quad H_1^{\text{rel}} : \rho(P_{\phi(Y)}, P_{\phi(X)}) > \rho(P_{\phi(Y)}, P_{\phi(Z)}). \tag{5}$$

This relative formulation removes the need to estimate the calibration offset ε. Instead of locating the valid band of a one-dimensional offset, TRIBET only requires the relative null condition that, when the online stream is HWT, $P_{\phi(Y)}$ is closer to the human reference $P_{\phi(X)}$ than to the machine reference $P_{\phi(Z)}$. In the same-domain setting, this condition is naturally satisfied when the testing human distribution matches the human reference distribution; in the cross-domain setting, it becomes an additional assumption. However, this assumption is more practical than estimating ε: it requires a qualitative relative closeness condition in the detector-score space, rather than an accurate source-only estimate of the testing-stream null mean. Empirically, Tab. 4 supports this point: on cross-domain failure cases where offset-based calibration can lead to FPRs up to 0.81, TRIBET keeps all FPRs below 0.05 across detector scores, while retaining strong power whenever the score itself is informative.

The rest of this section presents the two components of TRIBET. Section 4.1 constructs the triple-sample kernel-witness payoff and converts it into a conformal e-process. Section 4.2 then adapts this pipeline to the cross-domain setting through test-inclusive domain-shift correction.

4.1 From triple-sample payoff to conformal e-process

We instantiate the discrepancy ρ in (5) by the MMD distance [30] in the detector-score space. For sequential detection, however, we need a pointwise payoff for each incoming query rather than a batch-level MMD statistic. We therefore use the kernel witness [33] associated with this relative MMD. Let ℓ be a positive-definite kernel on the detector-score space; with the Gaussian RBF kernel used in our experiments, $0 \leq \ell \leq 1$. Given reference score sets $S_X = \{s_{x,i} = \phi(x_i)\}_{i=1}^m$ and $S_Z = \{s_{z,j} = \phi(z_j)\}_{j=1}^n$, we propose the *triple-sample kernel-witness payoff* for each query y_t as

$$w\left(\phi(y_t); S_X, S_Z\right) = \frac{1}{n}\sum_{j=1}^{n} \ell(\phi(y_t), s_{z,j}) - \frac{1}{m}\sum_{i=1}^{m} \ell(\phi(y_t), s_{x,i}). \tag{6}$$

Since $0 \leq \ell \leq 1$, we have $w(\phi(y_t); S_X, S_Z) \in [-1, 1]$.

The kernel ℓ measures similarity in the detector-score space: larger kernel values indicate higher similarity. Thus, smaller values of w indicate that the query is closer to the human reference, whereas larger values indicate stronger similarity to the machine reference. This payoff need not be an exact unbiased estimator of the full relative MMD, since it is computed from the finite reference samples $\mathcal{X}$ and $\mathcal{Z}$. Its role is to provide a bounded sequential evidence score that reveals relative human-machine similarity for each query. Besides, the payoff w provides a relative evidence score, but its absolute scale is not directly calibrated under the human null. We therefore follow the p-to-e route: first convert w into a valid conformal p-value by comparing it with calibration payoffs [34], and then calibrate this p-value into an e-value [35].

Conformal p-value. Let $\mathcal{C} = \{c_i\}_{i=1}^{n_{\text{cal}}}$ be a small calibration set (e.g., $n_{\text{cal}} \leq 20$) sampled from the target-domain human distribution. At each round t, we compare the payoff of the current query y_t with those of the calibration samples. Using the same reference score sets S_X and S_Z, we compute

$$w_{c,i} = w(\phi(c_i); S_X, S_Z), \qquad w_y = w(\phi(y_t); S_X, S_Z)\,.$$

Since larger values of w indicate stronger similarity to the machine reference, we use the right-tailed conformal p-value as follows

$$p_t = \frac{1 + \sum_{i=1}^{n_{\text{cal}}} \mathbf{1}\{w_{c,i} \geq w_y\}}{n_{\text{cal}} + 1}\,. \tag{7}$$

In the ideal case where the relative null H_0^{rel} is realized by a testing-stream human null, the samples $c_1, \ldots, c_{n_{\text{cal}}}, y_t$ are exchangeable conditional on $\mathcal{F}_{t-1}$. Since the same deterministic payoff is applied to every sample using the same reference score sets, the payoff values $w_{c,1}, \ldots, w_{c,n_{\text{cal}}}, w_y$ are also

exchangeable conditional on $\mathcal{F}_{t-1}$. Therefore, $\Pr_{H_0^{\mathrm{rel}}}(p_t \le u \mid \mathcal{F}_{t-1}) \le u,\ u \in [0,1]$. This conditional super-uniformity is the key input that allows us to convert the p-value into an e-value.

From p-values to e-values. We convert the conformal p-value into an e-value using calibrator:

$$e_t = \kappa_t p_t^{\kappa_t - 1}, \qquad \kappa_t \in [\kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}}] \subset (0,1), \tag{8}$$

where κ_t is predictable, i.e., chosen using only information available before observing p_t. Since p_t is conditionally super-uniform under H_0^{rel}, this calibration gives $\mathbb{E}_{H_0^{\mathrm{rel}}}[e_t \mid \mathcal{F}_{t-1}] \le 1$. The wealth process is then updated multiplicatively:

$$W_t = W_{t-1} e_t, \qquad W_0 = 1\ .$$

The test rejects H_0^{rel} at the first time t such that $W_t \ge 1/\alpha$. By Ville's inequality, this gives an anytime level-α test under the stated conditional super-uniformity condition.

Adaptive κ selection. The parameter κ_t controls the aggressiveness of the p-to-e calibration. It affects the wealth growth under H_1^{rel}, but not the false-positive guarantee, as long as κ_t is predictable. Following the log-growth view of e-values [35, 36], we choose κ_t by estimating the oracle constant $\kappa^\star$ that maximizes the expected log-growth $\mathbb{E}_{H_1^{\mathrm{rel}}}[\log(e_t)] = \mathbb{E}_{H_1^{\mathrm{rel}}}[\log(\kappa p_t^{\kappa-1})]$. Specifically, after a burn-in period t_0, we set

$$\kappa_t = \mathrm{clip}\left(\frac{1}{\widehat{M}_{t-1}}, \kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}}\right), \quad \text{with} \quad \widehat{M}_{t-1} = \frac{1}{t - t_0 - 1} \sum_{s=t_0+1}^{t-1} -\log p_s, \quad t > t_0 + 1.$$

The motivation is simple: if $M = \mathbb{E}_{H_1^{\mathrm{rel}}}[-\log p_t]$, then $\mathbb{E}_{H_1^{\mathrm{rel}}}[\log(\kappa p_t^{\kappa-1})] = \log\kappa + (1-\kappa)M$, whose unconstrained maximizer is $\kappa^\star = 1/M$. Thus, $\widehat{M}_{t-1}$ provides a predictable plug-in estimate of M, and κ_t estimates the clipped oracle choice without using the current p_t. The full streaming implementation of TRIBET with adaptive κ_t is given in Algorithm 1 in Appendix A.

The above validity argument relies on an ideal conditional exchangeability.

Assumption 1 (Round-wise exchangeable calibration)**.** *Under the ideal testing-stream HWT null that realizes H_0^{rel}, for every round t, the calibration samples used at that round and the current query y_t are exchangeable conditional on $\mathcal{F}_{t-1}$. Equivalently, the conformal p-value constructed from these samples is conditionally super-uniform:*

$$\Pr_{H_0^{\mathrm{rel}}}(p_t \le u \mid \mathcal{F}_{t-1}) \le u, \qquad u \in [0,1].$$

Under this ideal condition, Theorem 1 proves that TRIBET controls the anytime false-positive probability at level α. We next clarify the gap between the condition and practical implementation, which works well empirically.

Remark 1. *The formal guarantee requires conditional super-uniformity of p_t, which holds exactly when the calibration samples used at round t are fresh target-domain HWT samples exchangeable with y_t. In practice, fresh samples are usually unavailable. We therefore use a practical approximation: at each round, we draw a small mini-calibration set $\mathcal{C}_t$ from a larger target-domain human pool and compute the conformal rank using $\mathcal{C}_t \cup \{y_t\}$. This avoids reusing the same fixed small set throughout the stream, but it is a practical implementation rather than an exact conditional-exchangeability guarantee. Empirically, even under domain shift, $n_{\mathrm{cal}} = 5$ keeps the FPR below the nominal level while preserving strong power as shown in Tables 2 and 4.*

The construction above gives the basic payoff-to-e-process pipeline. We now adapt it to the cross-domain setting, where the reference scores may not be aligned with the testing stream.

4.2 TIDS: Test-Inclusive Domain-Shift Correction

The online-calibration dilemma in Section 3 shows that a source-only calibration offset ε is difficult to identify under domain shift. In TRIBET, we avoid estimating ε by replacing offset-based calibration with a relative comparison between human and machine references. The key challenge under domain shift is to preserve the relative *null condition*: when the online stream is HWT, its score distribution should remain closer to the human reference than to the machine reference.

To see why domain correction is needed, consider how target-domain HWT scores may move relative to the source references. If they become closer to the human reference and farther from the machine reference, the relative null condition is naturally preserved. In contrast, the other cases are challenging: HWT samples may become farther from the human reference and closer to the machine reference, making them appear spuriously machine-like and inflating false positives; or they may become closer to both references or farther from both references by different amounts. In the latter case, the relative closeness to the human and machine references can still change, making the relative human-machine comparison unreliable. Together, these challenging cases motivate our domain correction, which adjusts the reference scores based on the relative payoff.

A direct way to address this issue is to align the source-human reference scores to the target-domain human scores before computing the payoff. At round t, we estimate the *test-inclusive* score shift

$$\Delta_t = \frac{1}{n_{\text{cal}}+1}\left(\sum_{i=1}^{n_{\text{cal}}}\phi(c_i)+\phi(y_t)\right)-\frac{1}{m}\sum_{i=1}^{m}\phi(x_i).$$

In the ideal testing-stream HWT null, including y_t keeps $\mathcal{C}\cup\{y_t\}$ exchangeable when computing the conformal p-value. We then shift the source-human reference scores as

$$\widetilde{S}_X^{(t)} = \{\phi(x_i)+\Delta_t\}_{i=1}^{m}.$$

This shift increases the similarity between the target-domain human scores and the human reference in score space, while keeping the conformal comparison symmetric.

To further stabilize the relative comparison in the challenging cases above, we also apply the same shift to the machine reference:

$$\widetilde{S}_Z^{(t)} = \{\phi(z_j)+\Delta_t\}_{j=1}^{n}.$$

Intuitively, when target-domain HWT moves closer to the source machine reference, shifting only the human reference may still leave the machine anchor spuriously close to HWT samples. Applying the same shift to the machine reference moves both anchors to the target scale and carries the source human–machine separation into the corrected comparison. Likewise, in the uneven joint-shift case, combining a shifted human anchor with an unshifted machine anchor would mix target-aligned and source-domain references, making the relative comparison unreliable. Thus, the shared shift is a payoff-design choice that stabilizes the relative human–machine comparison under domain shift; conformal validity still comes from the test-inclusive symmetry.

Finally, at round t, we compute the corrected payoff for each candidate point $u\in\mathcal{C}\cup\{y_t\}$ as

$$w\left(\phi(u);\widetilde{S}_X^{(t)},\widetilde{S}_Z^{(t)}\right),$$

and then construct the conformal p-value and e-value as in Section 4.1. Thus, TIDS only replaces the original reference score sets S_X and S_Z in the payoff (6) with the shifted sets $\widetilde{S}_X^{(t)}$ and $\widetilde{S}_Z^{(t)}$.

Validity and power under shift. The shared translation is a power-design choice. For a translation-invariant kernel, an exact common location shift of both classes preserves source-domain kernel comparisons when references are shifted by the true translation. Class-dependent or nonlinear shifts need not preserve this alignment, and the estimated test-inclusive shift may be inaccurate. Null validity instead follows from round-wise conditional exchangeability and symmetric scoring; it does not require an exact translation model.

5 Experimental results & analysis

Experimental setup. We evaluate TRIBET in a black-box online MGT detection setting. Unless otherwise stated, the main results use Fast-DetectGPT [13] as the offline detector score, following the strongest detector setting in Chen and Wang [21]. To test score-function robustness, we also evaluate Binoculars [15], DetectGPT [12], and NPR [14]; these additional results are reported in Appendix D. Each detector score is computed using one of 8 open-weight proxy LLMs from the Llama [37], Falcon [38], Gemma [39, 40], GPT-Neo [41], and T5 [42] families; for each source and score function, we select the best proxy with the highest held-out AUROC.

Source model and data streams. We test on 15 source LLMs: 12 frontier models from the OpenAI [43], Google [44], and Anthropic [45] families, plus 3 pre-era models reused from Chen and Wang

Table 2: Main experiments on Olympic self-reference and XSum→Olympic cross-domain streams. Each row uses the proxy with the highest held-out Olympic AUROC $\hat{C}(f)$ for the corresponding source LLM. OOB-S1 denotes Chen and Wang [21] with an oracle ε; OOB-S2 uses the same payoff with an online warm-up estimate of the offset. TRIBET uses the triple-sample kernel-witness payoff with TIDS and adaptive online κ_t in §4. Each cell reports Pow ($\bar{\tau}$) / FPR over 200 test-stream permutations at level $\alpha = 0.05$ and warm-up samples $n_{\text{cal}} = 20$; † marks empirical FPR above α.

		Self-reference (Olympic→Olympic)			Cross-domain (XSum→Olympic)		
Source LLM	$\hat{C}(f)$	OOB-S1	OOB-S2	**TriBet**	OOB-S1	OOB-S2	**TriBet**
gpt4o-mini	0.94	1.00 (35) / 0.04	1.00 (81) / 0.01	**1.00 (4)** / 0.00	1.00 (77) / 0.01	1.00 (121) / 0.01	**1.00 (4)** / 0.01
gpt4o	0.90	1.00 (40) / 0.04	0.99 (100) / 0.01	**1.00 (5)** / 0.00	1.00 (114) / 0.01	1.00 (194) / 0.01	**1.00 (5)** / 0.01
gpt5.4	0.76	1.00 (72) / 0.04	0.30 (167) / 0.01	**1.00 (16)** / 0.00	0.01 (499) / 0.01	0.01 (499) / 0.02	**1.00 (13)** / 0.01
gpt5.4-pro	0.65	0.95 (141) / 0.02	0.02 (173) / 0.01	**1.00 (59)** / 0.01	0.01 (499) / 0.01	0.01 (499) / 0.01	**1.00 (59)** / 0.01
gemini-2.5-flash	0.71	1.00 (98) / **0.06**†	0.07 (172) / 0.01	**1.00 (28)** / 0.00	0.01 (499) / 0.02	0.01 (499) / 0.02	**1.00 (15)** / 0.00
gemini-2.5-pro	0.80	1.00 (58) / 0.05	0.50 (159) / 0.01	**1.00 (10)** / 0.00	0.01 (499) / 0.01	0.01 (499) / 0.01	**1.00 (10)** / 0.01
gemini-3.1-flash	0.82	1.00 (52) / **0.07**†	0.66 (150) / 0.01	**1.00 (9)** / 0.00	0.04 (499) / 0.01	0.03 (497) / 0.01	**1.00 (8)** / 0.01
gemini-3.1-pro	0.63	0.88 (150) / 0.02	0.02 (173) / 0.01	**1.00 (90)** / 0.00	0.01 (499) / 0.03	0.01 (499) / **0.20**†	**1.00 (100)** / 0.01
claude-opus-4.6	0.89	1.00 (38) / 0.04	0.99 (99) / 0.01	**1.00 (5)** / 0.00	1.00 (143) / 0.02	0.97 (254) / 0.01	**1.00 (5)** / 0.01
claude-sonnet-4.6	0.79	1.00 (58) / 0.05	0.53 (159) / 0.01	**1.00 (11)** / 0.00	0.02 (499) / 0.01	0.01 (499) / 0.01	**1.00 (9)** / 0.01
claude-opus-4.5	0.87	1.00 (44) / 0.04	0.95 (116) / 0.01	**1.00 (6)** / 0.00	1.00 (208) / 0.01	0.71 (373) / 0.01	**1.00 (6)** / 0.01
claude-haiku-4.5	0.81	1.00 (58) / 0.04	0.55 (159) / 0.01	**1.00 (11)** / 0.00	0.02 (499) / 0.01	0.01 (499) / 0.01	**1.00 (9)** / 0.01

[21]. Streams are drawn from three corpora: BBC XSum [28], our Olympic-2024 official news collection [29], and a long-context Student Essay corpus [46]. We consider two main regimes. In *self-reference*, the reference, calibration, and test streams are all drawn from the same stream. In *cross-domain*, the reference set is drawn from XSum (shortest) while the calibration and test streams are drawn from Olympic-2024 or Student Essay. Mixed-MGT streams, varying α, varying n_{cal}, and additional detector backbones can be referred to Appendix D.

Baselines, methods, and metrics. We compare TRIBET against the optimal online betting (OOB) baseline of Chen and Wang [21]. Baseline-S1 uses the two-sample plug-in payoff $|\phi(x_t) - \phi(y_t)|$ with an oracle null offset ε, while Baseline-S2 uses the same payoff but estimates ε from an online warmup. TRIBET replaces both components: the payoff becomes the triple-sample kernel witness $w(\phi(y); S_X, S_Z)$ in (6), and ONS-with-ε calibration is replaced by the TIDS conformal e-process in Section 4.2. The deployable version uses the adaptive online κ_t rule from Section 4.1; the optimal constant $\kappa^\star$ is used only as a diagnostic reference in Figure 2. We report power, mean stopping time $\bar{\tau}$, and FPR over repeated test-stream permutations at $\alpha = 0.05$. The two design moves, **(M1)** payoff replacement and **(M2)** strategy replacement, are isolated in the ablation study in Section 5.2.

Information available to each method. In the cross-domain runs, X is source-domain human text. OOB-S1 uses an oracle offset estimated from the target-human null, whereas OOB-S2 uses its warm-up offset protocol. TriBet uses target-human calibration data and an additional source-domain machine reference Z, generated offline by the detector's proxy rather than the audited source. Exact sample budgets and split protocols are reported in Appendix D.1.

5.1 Experimental Results

Power improvement. We first ask whether TRIBET improves online MGT detection in the regime that matters most for deployment: modern source LLMs, long streams, and possible mismatch between the reference corpus and the audited stream. Table 2 gives the main comparison on 12 modern source LLMs under both self-reference and cross-domain regimes. Across all 24 modern source–regime cells, TRIBET achieves 100% power while keeping FPR below the target level $\alpha = 0.05$. The improvement is not limited to the deployable OOB-S2 baseline, whose offset must be estimated online: TRIBET also outperforms OOB-S1, which is given oracle access to the null offset. This is most visible in the cross-domain regime. For several modern source LLMs, both OOB-S1 and OOB-S2 collapse to near-zero power, whereas TRIBET still rejects reliably within at most 100 rounds. Thus, the gain is not merely due to avoiding a noisy online estimate of ε; it comes from replacing the two-sample offset-based construction with target-calibrated relative evidence.

Stable FPR control across sources and detector scores. Across the modern-source results in Table 2, TRIBET keeps empirical FPR below the target level in every self-reference and cross-domain cell. Table 4 further shows that this stability is not specific to FastDetectGPT: the same TIDS calibration controls FPR for DetectGPT and NPR, including the heavy-tailed detector scores that are

challenging for ε-calibrated online betting. This supports the central design principle of TRIBET: false-positive control is inherited from conformal exchangeability with the calibration set, rather than from the tail behaviour or calibration of a particular detector score. Detector quality still matters for power and stopping time, as reflected by the held-out concordance $\widehat{C}(f)$, but the FPR guarantee remains stable across source models, stream regimes, and admissible detector scores.

Adaptive κ_t is near-optimal. Figure 2 explains the role of the adaptive κ_t schedule. Under H_1, AD closely tracks the optimal constant choice $\kappa^\star$ across strong, medium, and weak signals, while the fixed neutral choice $\kappa = 0.5$ can under-grow when the signal is too weak. Under H_0, all schedules remain below the rejection threshold. This confirms the theoretical separation: the conformal e-process maintain level-α validity, while κ_t controls how efficiently evidence is converted into stopping time.

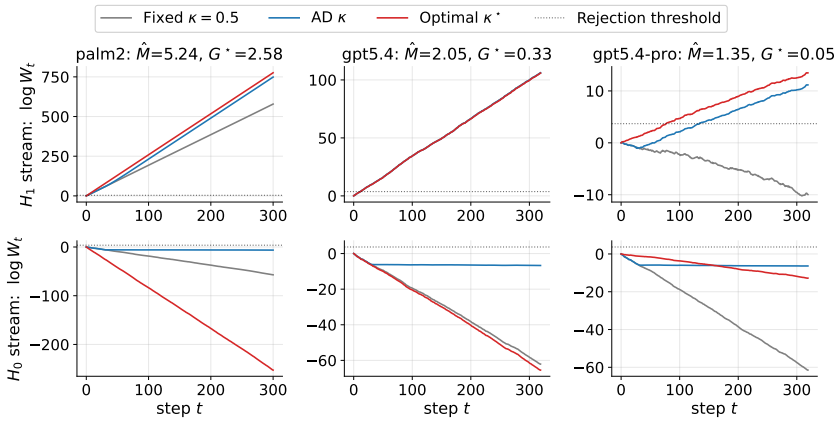


Figure 2: Average log-wealth under three predictable κ schedules: fixed neutral $\kappa = 0.5$, adaptive online AD, and the optimal constant choice $\kappa^\star$ computed in hindsight. Columns span strong, medium, and weak evidence regimes measured by $\widehat{M}$. Top row (H_1): AD tracks the optimal log-growth slope. Bottom row (H_0): all schedules remain below rejection threshold $\log(1/\alpha)$.

5.2 Ablation Study

We isolate the two design moves of TRIBET: **(M1)** replacing the two-sample payoff with the triple-sample kernel witness, and **(M2)** replacing ONS-with-ε calibration with the TIDS conformal e-process. Table 3 adds these components one at a time to the OOB baseline.

M1 improves the evidence signal. Holding the calibration mechanism fixed, replacing the two-sample payoff with the triple-sample witness improves power in both self-reference and cross-domain regimes. The gain is largest under cross-domain shift, where adding the machine reference $\mathcal{Z}$ recovers much of the power lost by the two-sample payoff. This supports the role of $\mathcal{Z}$ as a relative anchor: it does not provide validity by itself, but it gives the payoff more discriminative information.

M2 controls FPR and accelerates rejection. Adding TIDS on top of M1 removes the unstable FPR violations and further reduces rejection stopping time. This confirms that the TIDS component is not merely an implementation detail, but it is what converts the payoff into an anytime-valid e-process without estimating a fragile offset ε. Together, M1 and M2 implement the intended decoupling of TRIBET: the payoff supplies power, while TIDS supplies finite-sample FPR control.

6 Conclusion

We studied online machine-generated text detection as a sequential testing problem, where an audit system must accumulate evidence over a stream while controlling the probability of falsely accusing fully human-written content. We identified an *online-calibration dilemma* in existing two-sample betting approaches: the same offset that is needed to preserve false-positive control under the human null also determines whether the test has enough power to detect MGT. This coupling becomes fragile under testing-stream shift and detector-score misspecification. To address this, we proposed TRIBET, a triple-sample betting framework that decouples evidence construction from calibration. TRIBET uses both human and machine references to build a bounded relative kernel-witness payoff, and calibrates the resulting evidence through a test-inclusive domain-shift conformal e-process. This

Table 3: Component ablation of TRIBET. Rows add the two design moves sequentially: **(M1)** replaces the two-sample payoff with the triple-sample kernel witness, and **(M2)** replaces ONS-with-ε calibration with the TIDS conformal e-process. Each power cell reports average Pow and ($\bar{\tau}$), which is the mean stopping time; Power and FPR are averaged over the 30 corresponding source–mode cells. The "viol" column counts cells with empirical FPR exceeding $\alpha = 0.05$.

			Self-reference		**Cross-domain**		
Method	(M1)	(M2)	Pow ↑ ($\bar{\tau}$) ↓	FPR	Pow ↑ ($\bar{\tau}$) ↓	FPR	viol / 30
S1: oracle ε							
OOB-S1	×	×	0.969 (66)	0.037	0.542 (301)	0.013	2
OOB-S1+(M1)	✓	×	0.989 (39)	0.019	0.827 (68)	0.044	10
S2: online ε							
OOB-S2	×	×	0.622 (130)	0.013	0.510 (337)	0.032	2
OOB-S2+(M1)	✓	×	0.697 (129)	0.005	0.664 (128)	0.010	0
TriBet	✓	✓	**1.000 (20)**	**0.003**	**1.000 (36)**	**0.006**	**0**

yields anytime-valid false-positive control without estimating a null offset. Theoretically, TRIBET links power and stopping time to the rank separation induced by the payoff. Empirically, TRIBET consistently improves power and stopping time over online betting baselines across modern source LLMs, cross-domain streams, and challenging detector scores, while maintaining strict FPR control. These results suggest that relative evidence combined with conformal e-process calibration is a robust foundation for practical online MGT auditing.

Acknowledgments and Disclosure of Funding

This research was supported by the University of Melbourne's Research Computing Services and the Petascale Campus Initiative. XYT is supported by the Melbourne Research Scholarship and the ARC with grant number DE240101089 and Maincode AI Research Residency Program. ZJZ is supported by the Melbourne Research Scholarship and the ARC with grant number DE240101089. LHP is supported by the ARC with grant number LP240100101. Feng Liu is supported by the ARC with grant numbers DP230101540, DE240101089 and DP260103996.

References

[1] A. Yuan, A. Coenen, E. Reif, and D. Ippolito, "Wordcraft: Story writing with large language models," in *IUI*, 2022.

[2] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "ChatGPT for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.

[3] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *Transactions of the Association for Computational Linguistics*, vol. 12, 2024.

[4] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," in *NeurIPS*, 2019.

[5] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *ACL*, 2022.

[6] J. Lee, T. Le, J. Chen, and D. Lee, "Do language models plagiarize?" in *WWW*, 2023.

[7] C. Stokel-Walker, "AI bot ChatGPT writes smart essays — should professors worry?" *Nature*, 2022.

[8] D. R. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of chatgpt," *Innovations in education and teaching international*, vol. 61, no. 2, pp. 228–239, 2024.

[9] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv:2108.07258*, 2021.

[10] J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, and D. F. Wong, "A survey on LLM-generated text detection: Necessity, methods, and future directions," *Computational Linguistics*, vol. 51, no. 1, pp. 275–338, 2025.

[11] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," in *ICML*, 2023.

[12] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," in *ICML*, 2023.

[13] G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang, "Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature," in *ICLR*, 2024.

[14] J. Su, T. Zhuo, D. Wang, and P. Nakov, "DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text," in *EMNLP Findings*, 2023.

[15] 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," in *ICML*, 2024.

[16] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps *et al.*, "Release strategies and the social impacts of language models," *arXiv:1908.09203*, 2019.

[17] V. Verma, E. Fleisig, N. Tomlin, and D. Klein, "Ghostbuster: Detecting text ghostwritten by large language models," in *NAACL*, 2024.

[18] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can AI-generated text be reliably detected?" *arXiv:2303.11156*, 2023.

[19] N. Mireshghallah, J. Mattern, S. Gao, R. Shokri, and T. Berg-Kirkpatrick, "Smaller language models are better zero-shot machine-generated text detectors," in *EACL*, 2024.

[20] S. Chakraborty, A. S. Bedi, S. Zhu, B. An, D. Manocha, and F. Huang, "On the possibilities of AI-generated text detection," *arXiv:2304.04736*, 2023.

[21] C. Chen and J.-K. Wang, "Online detection of llm-generated texts via sequential hypothesis testing by betting," in *ICML*, 2025.

[22] G. Shafer, "Testing by betting: A strategy for statistical and scientific communication," *Journal of the Royal Statistical Society: Series A*, vol. 184, no. 2, pp. 407–431, 2021.

[23] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer, "Game-theoretic statistics and safe anytime-valid inference," *Statistical Science*, vol. 38, no. 4, pp. 576–601, 2023.

[24] I. Waudby-Smith and A. Ramdas, "Estimating means of bounded random variables by betting," *Journal of the Royal Statistical Society: Series B*, vol. 86, no. 1, pp. 1–27, 2024.

[25] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," *Machine Learning*, vol. 69, no. 2, pp. 169–192, 2007.

[26] A. Podkopaev and A. Ramdas, "Sequential predictive two-sample and independence testing," in *NeurIPS*, 2023.

[27] A. Saha and A. Ramdas, "Huber-robust likelihood ratio tests for composite nulls and alternatives," in *IEEE Transactions on Information Theory*, 2025.

[28] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," in *EMNLP*, 2018.

[29] Olympics, "Olympic games news," 2024. [Online]. Available: https://olympics.com/en/

[30] W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton, "A test of relative similarity for model selection in generative models," in *ICLR*, 2016.

[31] Y. Song, Z. Yuan, S. Zhang, Z. Fang, J. Yu, and F. Liu, "Deep kernel relative test for machine-generated text detection," in *ICLR*, 2025.

[32] A. Cutkosky and F. Orabona, "Black-box reductions for parameter-free online learning in banach spaces," in *COLT*, 2018.

[33] J. M. Kübler, W. Jitkrittum, B. Schölkopf, and K. Muandet, "A witness two-sample test," in *AISTATS*, 2022.

[34] G. Shafer and V. Vovk, "A tutorial on conformal prediction." *Journal of machine learning research*, vol. 9, no. 3, 2008.

[35] V. Vovk and R. Wang, "E-values: Calibration, combination and applications," *The Annals of Statistics*, vol. 49, no. 3, pp. 1736–1754, 2021.

[36] P. Grunwald, R. de Heide, and W. M. Koolen, "Safe testing," *Journal of the Royal Statistical Society: Series B*, vol. 86, no. 5, pp. 1091–1128, 2024, with discussion.

[37] Meta AI, "The Llama 3 herd of models," *arXiv:2407.21783*, 2024.

[38] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah *et al.*, "The Falcon series of open language models," *arXiv:2311.16867*, 2023.

[39] Gemma Team, Google DeepMind, "Gemma 2: Improving open language models at a practical size," *arXiv:2408.00118*, 2024.

[40] ——, "Gemma 3 technical report," *arXiv:2503.19786*, 2025.

[41] S. Black, L. Gao, P. Wang, C. Leahy, and S. Biderman, "GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow," Zenodo, 2021, https://doi.org/10.5281/zenodo.5297715.

[42] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[43] OpenAI, "GPT-4 technical report," *arXiv:2303.08774*, 2023.

[44] Gemini Team, Google, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv:2403.05530*, 2024.

[45] Anthropic, "The Claude 3 model family: Opus, Sonnet, Haiku," Anthropic Model Card, 2024, https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

[46] X. He, X. Shen, Z. Chen, M. Backes, and Y. Zhang, "Mgtbench: Benchmarking machine-generated text detection," in *CCS*, 2024.

[47] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.

[48] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[49] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[50] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *The Annals of Statistics*, vol. 51, no. 2, pp. 816–845, 2023.

A Algorithmic details

Algorithm 1: TRIBET Online MGT Detection with TIDS and Adaptive κ_t

Input: Human reference $\mathcal{X} = \{x_i\}_{i=1}^m$; machine reference $\mathcal{Z} = \{z_j\}_{j=1}^n$; known-human calibration set $\mathcal{C} = \{c_i\}_{i=1}^{n_{\text{cal}}}$; online query stream $(y_t)_{t\geq 1}$; level α; kernel k_σ; score function ϕ; burn-in t_0; bounds $0 < \kappa_{\text{lo}} \leq \kappa_{\text{hi}} < 1$.

Output: Rejection time τ if rejected, or $\perp$.

Compute reference scores $S_X \leftarrow \{\phi(x_i)\}_{i=1}^m$ and $S_Z \leftarrow \{\phi(z_j)\}_{j=1}^n$;

Compute $\bar{x} \leftarrow m^{-1}\sum_{i=1}^m \phi(x_i)$ and $S_C \leftarrow \sum_{i=1}^{n_{\text{cal}}} \phi(c_i)$;

Set $\log W \leftarrow 0$, $L \leftarrow \log(1/\alpha)$, and $\kappa_1 \leftarrow 0.5$;

for $t = 1, 2, \ldots$ **do**

  Observe y_t;

  Compute the test-inclusive shift $\Delta_t \leftarrow \frac{S_C + \phi(y_t)}{n_{\text{cal}}+1} - \bar{x}$.

  Shift the reference scores:

$$\widetilde{S}_X^{(t)} \leftarrow \widetilde{S}_X + \Delta_t, \qquad \widetilde{S}_Z^{(t)} \leftarrow \widetilde{S}_Z + \Delta_t.$$

  Compute the test payoff $w_y^{(t)} \leftarrow w(\phi(y_t); \widetilde{S}_X^{(t)}, \widetilde{S}_Z^{(t)})$ via (6);

  for $i = 1, \ldots, n_{\text{cal}}$ **do**

    Compute $w_{c,i}^{(t)} \leftarrow w(\phi(c_i); \widetilde{S}_X^{(t)}, \widetilde{S}_Z^{(t)})$;

  Compute the right-tailed conformal p-value

$$p_t \leftarrow \frac{1 + \sum_{i=1}^{n_{\text{cal}}} \mathbf{1}\{w_{c,i}^{(t)} \geq w_y^{(t)}\}}{n_{\text{cal}} + 1}.$$

  Convert to an e-value $e_t \leftarrow \kappa_t \cdot p_t^{\kappa_t - 1}$;

  Update wealth $\log W \leftarrow \log W + \log e_t$;

  if $\log W \geq L$ **then**

    return $\tau \leftarrow t$;

  if $t > t_0$ **then**

    Estimate

$$\widehat{M}_t \leftarrow \frac{1}{t - t_0} \sum_{s=t_0+1}^{t} -\log p_s.$$

    Set

$$\kappa_{t+1} \leftarrow \text{clip}\left(\frac{1}{\widehat{M}_t}, \kappa_{\text{lo}}, \kappa_{\text{hi}}\right).$$

  else

    Set $\kappa_{t+1} \leftarrow 0.5$;

return $\perp$;

B Theoretical Analysis

In this section, we summarize the guarantees of TRIBET, and the proofs and technical regularity conditions are given in Appendix C. The theory separates FPR validity from power: false-positive control follows from conformal exchangeability and predictable e-betting, while power and stopping time are dominated by the rank separation induced by the domain-shifted three-sample payoff.

Theorem 1 (Level-α test, power, and stopping time of TRIBET)**.** *Suppose Assumption 1 holds under the relative null H_0^{rel}. Then the TIDS conformal p-values are conditionally super-uniform:*

$$\Pr_{H_0^{\text{rel}}}(p_t \leq u \mid \mathcal{F}_{t-1}) \leq u, \qquad u \in [0, 1].$$

Then, for any predictable sequence $\kappa_t \in [\kappa_{\text{lo}}, \kappa_{\text{hi}}] \subset (0, 1)$, the wealth process

$$W_t = W_{t-1} \cdot \kappa_t p_t^{\kappa_t - 1}, \qquad W_0 = 1,$$

is a nonnegative supermartingale under H_0^{rel}. Consequently,

$$\Pr_{H_0^{\mathrm{rel}}}(\exists t \geq 1 : W_t \geq 1/\alpha) \leq \alpha.$$

Suppose Assumption 2 holds under the relative alternative H_1^{rel}, so that the conformal log-rank exponent $M = \mathbb{E}_{H_1^{\mathrm{rel}}}[-\log p_t]$ is well defined. For a fixed κ, the expected one-step log-growth is

$$G(\kappa; M) = \mathbb{E}_{H_1^{\mathrm{rel}}}[\log(\kappa p_t^{\kappa-1})] = \log \kappa + (1-\kappa)M.$$

The population-optimal constant aggressiveness is

$$\kappa^\star(M) = \mathrm{clip}\left(\frac{1}{M}, \kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}}\right).$$

If $M > 1$ and the alternative stream satisfies the regularity conditions in Appendix C, then the rejection time $\tau = \inf\{t : W_t \geq 1/\alpha\}$ is finite with probability one and, up to a bounded overshoot term B,

$$\mathbb{E}_{H_1^{\mathrm{rel}}}[\tau] \lesssim \frac{\log(1/\alpha) + B}{G(\kappa^\star; M)}.$$

Finally, let $\eta(\mathcal{Y}) = f(\mathcal{Y}; \mathcal{X}, \mathcal{Z})$ denote the payoff score and define the population concordance

$$C(f) := \Pr(\eta(\mathcal{Y}_h) < \eta(\mathcal{Y}_m)), \qquad Y_h \sim P_h, \quad Y_m \sim P_m.$$

Then

$$M \geq -\log(1 - C(f)).$$

Thus, stronger payoff concordance implies larger achievable log-wealth growth and shorter expected rejection time.

Theorem 1 has three implications. First, the false-positive guarantee is finite-sample and anytime-valid: it does not rely on a calibrated detector score, a null offset ϵ, or assumptions on the alternative. Second, the choice of payoff affects power but not validity; any admissible payoff that yields conditionally super-uniform conformal p-values preserves the level-α guarantee. Third, the expected stopping time is controlled by a rank-separation quantity M, which is lower-bounded by a monotone function of the payoff's population AUROC. This provides the theoretical link between offline separability and online detection speed.

C Theoretical Proofs

Throughout this appendix, we use the following shorthand. Let $\mathcal{C} = (c_1, \ldots, c_{n_{\mathrm{cal}}})$ denote the calibration set, let $N = n_{\mathrm{cal}} + 1$, and let $(\mathcal{F}_t)_{t \geq 0}$ be the filtration from the main text. Let

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i)$$

be the source-human reference score mean. For any tuple $U = (u_1, \ldots, u_N)$ of N points, define the test-inclusive score shift

$$\Delta(U) = \frac{1}{N} \sum_{j=1}^{N} \phi(u_j) - \bar{x}.$$

At round $t \geq 1$, we set $U_t = (c_1, \ldots, c_{n_{\mathrm{cal}}}, y_t)$ and $\Delta_t = \Delta(U_t)$. The shifted reference score sets are

$$\widetilde{S}_{X,t} = \{\phi(x_i) + \Delta_t\}_{i=1}^{m}, \qquad \widetilde{S}_{Z,t} = \{\phi(z_j) + \Delta_t\}_{j=1}^{n}.$$

For any admissible payoff f, define the round-t payoff of a candidate point $u \in \mathcal{C} \cup \{y_t\}$ as

$$\eta_t(u) = f(\phi(u); \widetilde{S}_{X,t}, \widetilde{S}_{Z,t}).$$

Thus, $w_y = \eta_t(y_t)$ and $w_{c,i} = \eta_t(c_i)$. Importantly, the shift Δ_t depends on the entire tuple U_t, not on the candidate point u being scored. Hence, within a fixed round t, all N candidate payoffs are computed using the same shifted reference score sets. This convention matches Algorithm 1 and is essential for Lemma 2 below.

C.1 Preliminaries: exchangeability and the Vovk calibrator

Definition 1 (Admissible payoff)**.** *A payoff function $f : \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, written as $f(r; S_X, S_Z)$, is* admissible *if it is deterministic, Borel-measurable, and permutation-invariant in the entries of each reference score set S_X and S_Z.*

Lemma 1 (Permutation invariance of the shift)**.** *The shift $\Delta(U)$ is a symmetric function of $U = (u_1, \ldots, u_N)$: for every permutation π of $\{1, \ldots, N\}$,*

$$\Delta(u_{\pi(1)}, \ldots, u_{\pi(N)}) = \Delta(u_1, \ldots, u_N).$$

Proof. By definition, $\Delta(U) = N^{-1} \sum_{j=1}^{N} \phi(u_j) - \bar{x}$. The sum is invariant to permutations of $(u_1, \ldots, u_N)$. □

Lemma 2 (Exchangeability of the payoffs under H_0^{rel}; cf. Vovk et al. 47)**.** *Fix $t \geq 1$. Suppose Assumption 1 holds under the relative null H_0^{rel}. Then, for any admissible payoff f, the tuple*

$$\big(\eta_t(c_1), \ldots, \eta_t(c_{n_{\mathrm{cal}}}), \eta_t(y_t)\big)$$

is exchangeable conditional on $\mathcal{F}_{t-1}$. Consequently, under a continuous payoff law or uniform random tie-breaking, the rank of $\eta_t(y_t)$ within this tuple is uniformly distributed on $\{1, \ldots, N\}$ conditional on $\mathcal{F}_{t-1}$.

Proof. By Assumption 1, under H_0^{rel}, the tuple $U_t = (c_1, \ldots, c_{n_{\mathrm{cal}}}, y_t)$ is exchangeable conditional on $\mathcal{F}_{t-1}$. By Lemma 1, the shift $\Delta_t = \Delta(U_t)$ is unchanged under any permutation of the entries of U_t. Hence the shifted reference score sets $\widetilde{S}_{X,t}$ and $\widetilde{S}_{Z,t}$ are also unchanged under such permutations.

For each candidate point $u \in \mathcal{C} \cup \{y_t\}$, the payoff is computed as

$$\eta_t(u) = f(\phi(u); \widetilde{S}_{X,t}, \widetilde{S}_{Z,t}).$$

Thus, permuting the entries of U_t only permutes the corresponding payoff values, because the same shifted reference score sets and the same admissible payoff are used for every candidate point. Therefore,

$$\big(\eta_t(c_1), \ldots, \eta_t(c_{n_{\mathrm{cal}}}), \eta_t(y_t)\big)$$

is exchangeable conditional on $\mathcal{F}_{t-1}$. The uniformity of the rank follows from exchangeability together with the continuity or random-tie-breaking convention. □

Lemma 3 (Vovk calibrator is an H_0^{rel}-valid e-value; Vovk and Wang 35)**.** *Let P be a $(0, 1]$-valued random variable with $\Pr(P \leq u) \leq u$ for all $u \in (0, 1]$. For any $\kappa \in (0, 1)$,*

$$\mathbb{E}[\kappa P^{\kappa - 1}] \leq 1.$$

Proof. Super-uniformity means that P stochastically dominates $U \sim \mathrm{Uniform}(0, 1]$. Equivalently, there exists a coupling such that $U \leq P$ almost surely. The function $g(p) = \kappa p^{\kappa - 1}$ is non-increasing on $(0, 1]$ for $\kappa \in (0, 1)$. Hence $g(P) \leq g(U)$ almost surely under this coupling. Taking expectations gives

$$\mathbb{E}[g(P)] \leq \mathbb{E}[g(U)] = \kappa \int_0^1 p^{\kappa - 1}\, dp = 1.$$

□

C.2 Propositions of Theorem 1

Proposition 1 (Anytime-valid level-α FPR)**.** *Suppose Assumption 1 holds. For any predictable $\kappa_t \in [\kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}}] \subseteq (0, 1)$, the wealth process*

$$W_t = W_{t-1} \kappa_t p_t^{\kappa_t - 1}, \qquad W_0 = 1,$$

is a nonnegative supermartingale under H_0^{rel}. Therefore, by Ville's inequality,

$$\Pr\left(\exists t \geq 1 : W_t \geq 1/\alpha \mid H_0^{\mathrm{rel}}\right) \leq \alpha.$$

Definition 2 (Conformal log-rank exponent and AUROC bridge)**.** *Let $\eta(Y) = f(\phi(Y); S_X, S_Z)$ denote the population payoff, and suppose that $\eta(Y_h)$ has a continuous distribution under the target-domain HWT distribution, with CDF F_h. Define*

$$p_\infty(y) = 1 - F_h(\eta(y)), \qquad M = \mathbb{E}_{Y_m}[-\log p_\infty(Y_m)],$$

where Y_m is drawn from the MGT alternative. We also define the concordance probability

$$C(f) = \Pr\big(\eta(Y_h) < \eta(Y_m)\big),$$

where Y_h and Y_m are independent HWT and MGT samples. By the Mann–Whitney identity [48, 49], the empirical AUROC consistently estimates $C(f)$. Moreover,

$$\mathbb{E}_{Y_m}[p_\infty(Y_m)] = 1 - C(f),$$

and Jensen's inequality gives

$$M \geq -\log\big(1 - C(f)\big). \tag{9}$$

We assume $M < \infty$ throughout.

Assumption 2 (Detectability)**.** *Under H_1^{rel}, the conformal log-rank exponent satisfies $M < \infty$, and the clipped oracle calibrator*

$$\kappa^\star(M) = \mathrm{clip}\left(\frac{1}{M}, \kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}}\right)$$

has positive expected log-growth:

$$G(\kappa^\star(M); M) > 0, \qquad G(\kappa; M) = \log\kappa + (1-\kappa)M.$$

In particular, if $1/M \in [\kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}}]$, this condition reduces to $M > 1$, since then

$$G(\kappa^\star(M); M) = M - 1 - \log M > 0.$$

Remark 2 (Verifiability of Assumption 2 from offline AUROC)**.** *By (9), $C(f) > 1 - 1/e \approx 0.632$ is sufficient for $M > 1$, and hence for positive oracle log-growth when $1/M$ lies in the allowed range of κ. This condition can be checked offline because $C(f)$ is consistently estimated by the empirical AUROC on a labeled held-out batch.*

Proposition 2 (Asymptotic power one)**.** *Suppose that, under H_1^{rel}, the chosen calibrator sequence satisfies*

$$\frac{1}{t}\sum_{s=1}^{t} \log e_s \to G > 0 \quad \textit{almost surely},$$

where $e_s = \kappa_s p_s^{\kappa_s - 1}$. Then

$$\Pr(\tau < \infty \mid H_1^{\mathrm{rel}}) = 1, \qquad \tau = \inf\{t : W_t \geq 1/\alpha\}.$$

This condition holds for the oracle fixed choice $\kappa^\star(M)$ when its expected log-growth is positive, and for the adaptive rule whenever its plug-in estimate converges to $\kappa^\star(M)$ and the corresponding empirical log-growth converges.

Proposition 3 (Expected rejection time and asymptotic GRO rate)**.** *Under the conditions of Proposition 2, suppose that under H_1^{rel} there exists a $\mathcal{C}$-measurable constant $c(\mathcal{C}) > 0$ such that, before rejection,*

$$c(\mathcal{C}) \leq \mathbb{E}_{H_1^{\mathrm{rel}}}[\log e_t \mid \mathcal{F}_{t-1}]. \tag{10}$$

Then the following statements hold.

(a) ***Conditional bound.*** *If $c(\mathcal{C}) > 0$, then*

$$\mathbb{E}[\tau \mid H_1^{\mathrm{rel}}, \mathcal{C}] \leq \frac{\log(1/\alpha) + B}{c(\mathcal{C})}, \qquad B = |\log\kappa_{\mathrm{lo}}| + (1 - \kappa_{\mathrm{lo}})\log N. \tag{11}$$

Here B is a deterministic one-step bound, using the fact that the conformal p-value is bounded below by $1/N$. If $\mathbb{E}[1/c(\mathcal{C})] < \infty$, marginalizing yields $\mathbb{E}[\tau \mid H_1^{\mathrm{rel}}] = O(\log(1/\alpha)/c^\star)$ with $c^\star = 1/\mathbb{E}[1/c(\mathcal{C})]$.

(b) ***Asymptotic GRO rate.*** *For a fixed κ, the expected one-step log-growth is*

$$G(\kappa; M) = \mathbb{E}_{H_1^{\mathrm{rel}}}[\log(\kappa p_t^{\kappa-1})] = \log\kappa + (1-\kappa)M.$$

The oracle choice within the Vovk family is

$$\kappa^\star(M) = \mathrm{clip}\left(\frac{1}{M}, \kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}}\right).$$

When $1/M \in [\kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}}]$, the optimal growth rate is

$$G^\star(M) = M - 1 - \log M.$$

The adaptive rule attains the same asymptotic rate whenever its plug-in estimate of M converges to the true M.

(c) ***Concordance-bounded asymptotic rate.*** *Combining (a)–(b) with (9), when $C(f) > 1 - 1/e$,*

$$\mathbb{E}[\tau \mid H_1^{\mathrm{rel}}] \lesssim \frac{\log(1/\alpha) + B}{m^\star - 1 - \log m^\star}, \qquad m^\star = -\log\big(1 - C(f)\big), \tag{12}$$

which decreases as the concordance $C(f)$ increases on this domain.

C.3 Proof of Proposition 1 (anytime-valid FPR)

Proof of Proposition 1. We show that $(W_t)_{t\geq 0}$ is a non-negative supermartingale under H_0 with respect to $(\mathcal{F}_t)_{t\geq 0}$; the conclusion then follows from Ville's inequality.

(i) Non-negativity. From (8) with $\kappa \in (0,1)$ and $p_t \in (0,1]$, $e_t = \kappa p_t^{\kappa-1} \geq \kappa > 0$, hence $W_t = W_{t-1} \cdot e_t \geq 0$ and $W_0 = 1$.

(ii) Supermartingale property under H_0. Fix $t \geq 1$. By Lemma 2, under H_0 the rank R_t of $\eta_t(y_t)$ within $\big(\eta_t(c_1), \ldots, \eta_t(c_{n_{\mathrm{cal}}}), \eta_t(y_t)\big)$ is uniform on $\{1, \ldots, N\}$ conditional on $\mathcal{F}_{t-1}$ (note that $y_t \notin \mathcal{F}_{t-1}$, so conditioning on $\mathcal{F}_{t-1}$ fixes only the calibration set and past queries). The conformal p-value in (7) can be written

$$p_t \;=\; \frac{\big|\{i :\, w_{c,i} \geq w_y\}\big| + 1}{N} \;=\; \frac{R_t^\star}{N},$$

where $R_t^\star \in \{1, \ldots, N\}$ is the (ties-adjusted) rank of w_y from the top.[3] Since $R_t^\star \sim \mathrm{Unif}\{1, \ldots, N\}$, p_t is super-uniform on $(0, 1]$. By Lemma 3,

$$\mathbb{E}\big[e_t \,\big|\, \mathcal{F}_{t-1}\big] \;=\; \mathbb{E}\big[\kappa\, p_t^{\kappa-1} \,\big|\, \mathcal{F}_{t-1}\big] \;\leq\; 1.$$

Consequently, under H_0,

$$\mathbb{E}\big[W_t \,\big|\, \mathcal{F}_{t-1}\big] \;=\; W_{t-1} \cdot \mathbb{E}\big[e_t \,\big|\, \mathcal{F}_{t-1}\big] \;\leq\; W_{t-1},$$

establishing the supermartingale property.

(iii) Ville's inequality. Since (W_t) is a non-negative supermartingale with $\mathbb{E}[W_0] = 1$, Ville's inequality gives, for any $t^\star \in \mathbb{N} \cup \{\infty\}$,

$$\Pr\left(\sup_{1\leq t\leq t^\star} W_t \geq 1/\alpha \,\middle|\, H_0^{\mathrm{rel}}\right) \;\leq\; \alpha\mathbb{E}[W_0] \;=\; \alpha.$$

Letting $t^\star \to \infty$ and noting the event is monotone in $t^\star$ yields the anytime-valid bound $\Pr(\exists t \geq 1 : W_t \geq 1/\alpha \mid H_0^{\mathrm{rel}}) \leq \alpha$.

The argument uses only: (a) Lemma 1 (which depends on the shift being a symmetric statistic of $\mathcal{C} \cup \{y_t\}$), (b) the composability condition on f (permutation-invariance in X and in Z, and determinism in (y, X, Z)), and (c) Lemma 3. It does *not* use any property of the null distribution P_h, nor any calibration of a nuisance ε. Hence the bound holds for any $n_{\mathrm{cal}} \geq 1$ and any composable f. □

[3] If ties occur, break them uniformly at random; the distribution of p_t is then exactly uniform on $\{1/N, 2/N, \ldots, 1\}$.

C.4 Proof of Proposition 2 (asymptotic power)

Proof of Proposition 2. Recall that the wealth process is updated multiplicatively as

$$W_t = \prod_{s=1}^{t} e_s.$$

Taking logarithms gives the cumulative log-wealth

$$S_t = \log W_t = \sum_{s=1}^{t} \log e_s.$$

The rejection rule stops at

$$\tau = \inf\{t : W_t \geq 1/\alpha\}.$$

Equivalently, since the logarithm is increasing, this is the first time that

$$S_t \geq L_\alpha, \qquad L_\alpha = \log(1/\alpha).$$

By assumption, under H_1^{rel},

$$\frac{S_t}{t} = \frac{1}{t}\sum_{s=1}^{t} \log e_s \to G > 0 \quad \text{almost surely.}$$

Therefore, on an event with probability one, there exists a finite random time T such that for all $t \geq T$,

$$\frac{S_t}{t} \geq \frac{G}{2}.$$

Here $G/2$ is only used as a positive margin below the limiting growth rate G. Hence, for all $t \geq T$,

$$S_t \geq \frac{G}{2} t.$$

Since $G/2 > 0$, the right-hand side goes to $+\infty$ as $t \to \infty$. Thus $S_t \to +\infty$ almost surely.

Because the rejection threshold $L_\alpha = \log(1/\alpha)$ is finite, there must exist a finite time t such that

$$S_t \geq L_\alpha.$$

Equivalently, $W_t \geq 1/\alpha$. Therefore the stopping time τ is finite almost surely under H_1^{rel}, namely

$$\Pr(\tau < \infty \mid H_1^{\mathrm{rel}}) = 1.$$

Finally, the oracle fixed choice and the adaptive rule are examples of calibrator choices for which the premise can hold. For the oracle fixed choice, if the empirical average log-growth converges to its expectation and

$$G(\kappa^\star(M); M) > 0,$$

then the required condition holds with limiting growth rate $G(\kappa^\star(M); M)$. For the adaptive rule, if the plug-in estimate of M converges to M, then κ_t converges to the clipped oracle choice $\kappa^\star(M)$. If the corresponding empirical average log-growth also converges, then the adaptive rule has the same positive limiting log-growth, and the same argument above applies. □

C.5 Proof of Proposition 3

Proof of Proposition 3. Let

$$L_t = \log e_t, \qquad S_t = \log W_t = \sum_{s=1}^{t} L_s, \qquad L_\alpha = \log(1/\alpha).$$

Then the rejection time can be written as

$$\tau = \inf\{t : S_t \geq L_\alpha\}.$$

Step 1: expected stopping-time bound. Since $p_t \in \{1/N, 2/N, \dots, 1\}$ and $\kappa_t \in [\kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}}] \subset (0,1)$, we have

$$L_t = \log \kappa_t + (\kappa_t - 1) \log p_t = \log \kappa_t + (1 - \kappa_t)(-\log p_t).$$

Hence L_t is uniformly bounded. In particular,

$$L_t \le B, \qquad B = |\log \kappa_{\mathrm{lo}}| + (1 - \kappa_{\mathrm{lo}}) \log N.$$

Condition on a realization of $\mathcal{C}$ with $c(\mathcal{C}) > 0$. By (10), before rejection,

$$\mathbb{E}_{H_1^{\mathrm{rel}}}[L_t \mid \mathcal{F}_{t-1}] \ge c(\mathcal{C}).$$

Therefore,

$$S_{t \wedge \tau} - c(\mathcal{C})(t \wedge \tau)$$

is a submartingale. Taking expectations gives

$$\mathbb{E}[S_{t \wedge \tau} \mid H_1^{\mathrm{rel}}, \mathcal{C}] \ge c(\mathcal{C}) \mathbb{E}[t \wedge \tau \mid H_1^{\mathrm{rel}}, \mathcal{C}].$$

On the other hand, by definition of τ, if $\tau \le t$ then $S_{\tau - 1} < L_\alpha$ and

$$S_\tau = S_{\tau - 1} + L_\tau \le L_\alpha + B.$$

If $\tau > t$, then $S_t < L_\alpha \le L_\alpha + B$. Hence, in all cases,

$$S_{t \wedge \tau} \le L_\alpha + B.$$

Combining the two displays,

$$c(\mathcal{C}) \mathbb{E}[t \wedge \tau \mid H_1^{\mathrm{rel}}, \mathcal{C}] \le L_\alpha + B.$$

Letting $t \to \infty$ and using monotone convergence yields

$$\mathbb{E}[\tau \mid H_1^{\mathrm{rel}}, \mathcal{C}] \le \frac{\log(1/\alpha) + B}{c(\mathcal{C})}.$$

This proves (11). If $\mathbb{E}[1/c(\mathcal{C})] < \infty$, taking expectation over $\mathcal{C}$ gives

$$\mathbb{E}[\tau \mid H_1^{\mathrm{rel}}] \le (\log(1/\alpha) + B) \mathbb{E}\left[\frac{1}{c(\mathcal{C})}\right],$$

which is the stated marginal bound with $c^\star = 1/\mathbb{E}[1/c(\mathcal{C})]$.

Step 2: asymptotic GRO rate. For a fixed $\kappa \in (0,1)$, the Vovk calibrator gives

$$e_t = \kappa p_t^{\kappa - 1}.$$

Thus,

$$\log e_t = \log \kappa + (\kappa - 1) \log p_t = \log \kappa + (1 - \kappa)(-\log p_t).$$

Taking expectation under H_1^{rel} and using

$$M = \mathbb{E}_{H_1^{\mathrm{rel}}}[-\log p_t],$$

we obtain

$$G(\kappa; M) = \mathbb{E}_{H_1^{\mathrm{rel}}}[\log e_t] = \log \kappa + (1 - \kappa) M.$$

Differentiating with respect to κ gives

$$\frac{\partial}{\partial \kappa} G(\kappa; M) = \frac{1}{\kappa} - M, \qquad \frac{\partial^2}{\partial \kappa^2} G(\kappa; M) = -\frac{1}{\kappa^2} < 0.$$

Therefore, the unconstrained maximizer is $\kappa = 1/M$. Restricting κ to $[\kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}}]$ gives

$$\kappa^\star(M) = \mathrm{clip}\left(\frac{1}{M}, \kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}}\right).$$

When $1/M \in [\kappa_{\mathrm{lo}}, \kappa_{\mathrm{hi}}]$, substituting $\kappa = 1/M$ gives

$$G^\star(M) = G(1/M; M) = M - 1 - \log M.$$

Table 4: Score-function robustness on the cross-domain failure cases from Chen and Wang [21]. All rows use the Gemma-2B proxy and Scenario 2 for OOB. Each cell reports power, mean stopping time $\bar{\tau}$, and FPR over 200 trials at $\alpha = 0.05$ and $n_{\text{cal}} = 20$. Marker † denotes empirical FPR above α. TRIBET controls FPR across all detector scores; bold power values mark valid head-to-head comparisons where both methods control FPR.

Source	Score	$\hat{C}(f)$	OOB Pow	OOB FPR	TriBet Pow	TriBet FPR
PaLM 2	FDGPT	0.999	1.00 (47)	**0.07**†	1.00 (5)	0.03
PaLM 2	DetGPT	0.838	1.00 (70)	**0.40**†	0.99 (20)	0.03
PaLM 2	NPR	0.914	1.00 (57)	**0.77**†	1.00 (10)	0.03
1.5-Flash	FDGPT	0.978	1.00 (71)	0.01	**1.00 (6)**	0.03
1.5-Flash	DetGPT	0.766	1.00 (84)	**0.60**†	0.93 (54)	0.04
1.5-Flash	NPR	0.786	1.00 (77)	**0.81**†	0.96 (42)	0.02
1.5-Pro	FDGPT	0.870	0.73 (377)	0.02	**1.00 (12)**	0.03
1.5-Pro	DetGPT	0.704	1.00 (101)	**0.65**†	0.72 (100)	0.04
1.5-Pro	NPR	0.705	1.00 (99)	**0.81**†	0.61 (122)	0.04

For the adaptive rule, if the plug-in estimate of M converges to the true M, then by continuity of the clipping map,

$$\kappa_t = \text{clip}\left(\frac{1}{\widehat{M}_t}, \kappa_{\text{lo}}, \kappa_{\text{hi}}\right) \to \kappa^\star(M).$$

Thus, whenever the empirical average log-growth converges along this sequence, the adaptive rule attains the same asymptotic GRO rate as the oracle choice.

Step 3: concordance-bounded asymptotic rate. By (9),

$$M \geq m^\star, \qquad m^\star = -\log\big(1 - C(f)\big).$$

When $C(f) > 1 - 1/e$, we have $m^\star > 1$. On the region $m > 1$, the function

$$g(m) = m - 1 - \log m$$

is increasing because

$$g'(m) = 1 - \frac{1}{m} > 0.$$

Therefore,

$$G^\star(M) = M - 1 - \log M \geq m^\star - 1 - \log m^\star.$$

Combining this lower bound on the asymptotic log-growth with the stopping-time bound in Step 1 gives

$$\mathbb{E}[\tau \mid H_1^{\text{rel}}] \lesssim \frac{\log(1/\alpha) + B}{m^\star - 1 - \log m^\star}, \qquad m^\star = -\log\big(1 - C(f)\big).$$

Since $g(m)$ is increasing for $m > 1$ and $m^\star$ increases with $C(f)$, the bound decreases as the concordance $C(f)$ increases. This proves (12). □

D Extra experimental results and implementation details

This appendix provides implementation details and robustness checks supporting Section 5. Section D.1 describes the corpora, source LLMs, generation protocols, proxy models, and detector-score computation. Section D.2 reports additional results on the Student Essay corpus. Section D.3 defines the detector scores used throughout the paper. Sections D.4, D.5, and D.6 report robustness checks on alternative detector scores, mixed human–machine streams, and sensitivity to the rejection level α, respectively. Unless otherwise stated, TRIBET refers to the deployable adaptive-κ_t variant.

D.1 Data generation and corpora

Our evaluation extends the protocol of Chen and Wang [21] along three axes: newer source LLMs, a post-cutoff Olympic-2024 news corpus, and a long-context Student Essay corpus. Table 5 summarizes the resulting per-source word-count statistics after generation and length normalization.

Corpora. We use three human-written corpora, each paired with machine-generated text from the source LLMs.

- **XSum** [28]: a BBC-news summarization corpus containing approximately 228,000 articles. We use the same 500-article subset as Chen and Wang [21], with mean length 116 words and range 26–165. XSum serves as the reference corpus in the cross-domain setting.
- **Olympic-2024** [29]: a collection of 519 news articles covering the Paris 2024 Summer Olympics. The articles were published in July–August 2024, after the pretraining cutoffs of the proxy LLMs used for scoring, reducing the risk that proxy scores are affected by memorization of the source articles. The raw articles have mean length 217 words and median length 197 words. After the length-normalization procedure described below, paired human and machine texts have mean length approximately 139 words.
- **Student Essay** [46]: a long-context corpus of 1,000 student essays, each paired with its original assignment prompt. We cap essays at 650 words, yielding a substantially longer stream regime than the news corpora. This setting tests whether online detection remains effective when evidence is accumulated across manuscript-like documents rather than short news spans.

Source LLMs. We evaluate 12 modern source LLMs: four OpenAI models (GPT-4o-mini, GPT-4o, GPT-5.4, GPT-5.4-pro), four Google models (Gemini-2.5-Flash, Gemini-2.5-Pro, Gemini-3.1-Flash, Gemini-3.1-Pro), and four Anthropic models (Claude-Opus-4.6, Claude-Sonnet-4.6, Claude-Opus-4.5, Claude-Haiku-4.5). For backward comparability with Chen and Wang [21], we also reuse three paper-era machine corpora: PaLM 2, Gemini-1.5-Flash, and Gemini-1.5-Pro.

Machine-text generation. We use two generation protocols, depending on the corpus.

For **news corpora** (XSum and Olympic-2024), we follow the prefix-continuation protocol of Chen and Wang [21]. For each human article, the first 30 tokens are extracted as a prefix and submitted to the source LLM with the prompt:

> *"You are a news writer. Please write an article with about* 150 *words starting exactly with* ⟨*prefix*⟩*."*

Sampling uses temperature 1.0 and `max_new_tokens`= 300. To control for length effects, the paired human article is truncated to match the generated machine text length. Thus, each stored news pair contains a human and machine span of comparable length before detector scoring.

For **Student Essay**, we use prompt-only generation. The assignment prompt is submitted to the source LLM without any human prefix, using temperature 1.0 and `max_new_tokens`= 1500. Human and machine essays are then truncated to 650 words for length parity.

For **mixed-MGT streams**, we use Olympic-2024 articles and vary the machine-generated fraction $p \in \{5, 10, 20, 30, 50, 70, 100\}\%$. The first $1-p$ fraction of each article is retained as human text, and the source LLM completes the remaining portion to the original article length. The case $p = 100\%$ reduces to the standard fully machine-generated setting.

Failure handling and filtering. Generations that fail because of API timeouts, content filters, or outputs shorter than 10 tokens are retried up to five times. If all retries fail, we use a simpler continuation prompt:

> *"Continue writing the following text to about* 150 *words total . . . "*

Persistent failures are marked explicitly as `[GENERATION_FAILED]` and are not replaced by human text, to avoid contaminating the machine distribution.

Machine reference sets. We distinguish between two kinds of machine-generated text. The *audited MGT streams* are generated by the source LLMs listed above and are used only for evaluation.

In contrast, the machine reference set $\mathcal{Z}$ used by TRIBET is generated offline by the proxy LLM associated with the detector score ϕ. Thus, for each proxy LLM, both the detector score function ϕ and the machine reference scores $\{\phi(z) : z \in \mathcal{Z}\}$ are defined using the same proxy model. The audited source LLM is treated as unknown and is never used to generate $\mathcal{Z}$.

Proxy LLMs and detector scoring. Each human–machine text pair is scored offline using proxy LLMs from the Falcon, Gemma, GPT-Neo, Llama, and T5 families:

$$\{\text{Falcon-7B},\ \text{Falcon-40B},\ \text{Gemma-2B},\ \text{Gemma-3-4B},\ \text{Gemma-3-12B},\ \text{GPT-Neo-2.7B},\ \text{Llama-3.1-8B},\ \text{Llama-3.2-3B}\}.$$

The main experiments use Fast-DetectGPT [13]; additional robustness experiments use DetectGPT [12], NPR [14], and Binoculars [15] as described in Section D.3. For headline tables, the proxy is selected separately for each source by maximizing held-out AUROC on Olympic-2024 under the corresponding detector score.

Discussion of the feasibility of the testing-stream calibration set. TRIBET requires a small set $\mathcal{C}$ of confirmed-human calibration samples representative of the audit stream. We view this requirement as realistic in many operational settings. News organisations have archived pre-LLM articles, educational institutions maintain historical student submissions, academic venues retain accepted manuscripts, and content platforms often have human-authenticated text collected before the deployment of capable LLMs. These sources typically provide far more than the $n_{\text{cal}} = 20$–200 samples used in our experiments. The role of $\mathcal{C}$ is also lighter than that of a supervised training set: it is used only for rank calibration, not for fitting a classifier. Smaller n_{cal} gives coarser conformal p-values with resolution $1/(n_{\text{cal}} + 1)$, while larger n_{cal} improves calibration resolution and power. If the audit stream drifts over time, the calibration set can be refreshed with newly verified human samples without retraining the detector; extending the finite-sample guarantee to explicit distribution drift is closely related to beyond-exchangeability conformal prediction [50].

Table 5: Per-source word-count statistics after generation and length normalization. Each cell reports n/mean/median/min/max word counts for machine-generated texts. For XSum and Olympic-2024, the paired human text is truncated to the generated machine length, so the stored human and machine spans are length-matched. The final row reports raw human corpus statistics before per-source pairing and truncation.

Source	XSum (n / mean / median / min / max)	Olympic-2024 (n / mean / median / min / max)	Student Essay (n / mean / median / min / max)
gpt4o-mini	500 / 116 / 118 / 26 / 159	519 / 137 / 150 / 44 / 166	1000 / 646 / 650 / 440 / 650
gpt4o	500 / 116 / 118 / 26 / 161	519 / 139 / 153 / 44 / 169	1000 / 637 / 650 / 532 / 650
gpt5.4	500 / 116 / 118 / 26 / 168	519 / 140 / 154 / 44 / 174	1000 / 650 / 650 / 481 / 650
gpt5.4-pro	500 / 116 / 118 / 26 / 173	519 / 137 / 149 / 44 / 194	1000 / 649 / 650 / 575 / 650
gemini-2.5-flash	500 / 115 / 118 / 26 / 174	519 / 134 / 141 / 44 / 194	1000 / 650 / 650 / 580 / 650
gemini-2.5-pro	500 / 115 / 118 / 26 / 156	519 / 132 / 142 / 44 / 155	1000 / 649 / 650 / 583 / 650
gemini-3.1-flash	500 / 116 / 118 / 26 / 156	519 / 137 / 152 / 44 / 160	1000 / 632 / 636 / 557 / 650
gemini-3.1-pro	500 / 120 / 118 / 100 / 161	519 / 142 / 151 / 100 / 178	1000 / 593 / 617 / 228 / 650
claude-opus-4.6	500 / 116 / 118 / 26 / 165	519 / 139 / 152 / 44 / 188	1000 / 648 / 650 / 1 / 650
claude-sonnet-4.6	500 / 116 / 118 / 26 / 164	519 / 139 / 153 / 44 / 173	1000 / 648 / 650 / 1 / 650
claude-opus-4.5	500 / 116 / 118 / 26 / 169	519 / 140 / 154 / 44 / 171	1000 / 645 / 650 / 1 / 650
claude-haiku-4.5	500 / 116 / 118 / 26 / 165	519 / 138 / 149 / 44 / 176	1000 / 644 / 650 / 578 / 650
Human (raw)	500 / 116 / 118 / 26 / 165	519 / 217 / 197 / 44 / 688	1000 / 656 / 530 / 203 / 6109

D.2 Result on the Student Essay dataset

We next evaluate a long-context setting using the Student Essay corpus [46]. Each stream is a single essay split into at most $T_{\max} = 15$ contiguous chunks, mimicking a document-level audit in which evidence arrives paragraph by paragraph. We use Falcon-7B as the proxy and FastDetectGPT as the detector score for all 12 modern source LLMs.

The Student Essay corpus is more discriminable than Olympic under Fast-DetectGPT with Falcon-7B: $\hat{C}(f) \geq 0.93$ on each of the 12 source LLMs, since essay-length spans afford the proxy substantially more context than news leads. TRIBET converts this AUROC headroom into stopping times $\bar{\tau} \leq 9$ chunks on every cell, while OOB does not exceed 43% power, because the $T_{\max} = 15$ stream cap is

Table 6: Student Essay self-reference and XSum→Essay cross-domain results. Each cell reports Pow $(\bar{\tau})$/FPR over 200 essay-level trials at $\alpha = 0.05$, with $T_{\max} = 15$ chunks and $n_{\text{cal}} = 20$. $\hat{C}(f)$ is the held-out Essay AUROC of FastDetectGPT under Falcon-7B. Marker † denotes empirical FPR above α.

		Self-reference (Essay→Essay)			Cross-domain (XSum→Essay)		
Source LLM	$\hat{C}(f)$	OOB-S1	OOB-S2	**TriBet**	OOB-S1	OOB-S2	**TriBet**
gpt4o-mini	1.00	0.28 (15) / 0.02	0.01 (15) / 0.02	**1.00 (5)** / 0.00	0.32 (15) / 0.03	0.01 (15) / 0.01	**1.00 (8)** / 0.00
gpt4o	1.00	0.20 (15) / 0.02	0.03 (15) / 0.01	**1.00 (5)** / 0.00	0.28 (15) / 0.01	0.03 (15) / 0.04	**1.00 (7)** / 0.00
gpt5.4	0.97	0.11 (15) / 0.04	0.03 (15) / 0.01	**1.00 (6)** / 0.00	0.11 (15) / 0.01	0.05 (15) / **0.06**†	**1.00 (6)** / 0.00
gpt5.4-pro	0.93	0.10 (15) / 0.04	0.03 (15) / 0.04	**1.00 (7)** / 0.00	0.07 (15) / 0.04	0.04 (15) / 0.03	**1.00 (7)** / 0.00
gemini-2.5-flash	1.00	0.33 (15) / 0.04	0.04 (15) / 0.03	**1.00 (5)** / 0.00	0.35 (15) / 0.01	0.02 (15) / 0.04	**1.00 (8)** / 0.00
gemini-2.5-pro	1.00	0.36 (15) / 0.02	0.01 (15) / 0.04	**1.00 (5)** / 0.00	0.43 (15) / 0.01	0.01 (15) / 0.01	**1.00 (8)** / 0.00
gemini-3.1-flash	1.00	0.33 (15) / 0.01	0.04 (15) / 0.03	**1.00 (5)** / 0.00	0.30 (15) / 0.01	0.04 (15) / 0.03	**1.00 (8)** / 0.00
gemini-3.1-pro	0.97	0.10 (15) / **0.06**†	0.02 (15) / 0.02	**1.00 (6)** / 0.00	0.17 (15) / 0.01	0.03 (15) / 0.00	**1.00 (6)** / 0.00
claude-opus-4.6	1.00	0.20 (15) / 0.01	0.04 (15) / 0.02	**1.00 (5)** / 0.00	0.30 (15) / 0.01	0.04 (15) / 0.03	**1.00 (7)** / 0.00
claude-sonnet-4.6	0.99	0.14 (15) / 0.04	0.04 (15) / 0.04	**1.00 (5)** / 0.00	0.18 (15) / 0.02	0.05 (15) / 0.04	**1.00 (5)** / 0.00
claude-opus-4.5	1.00	0.21 (15) / 0.03	0.04 (15) / 0.01	**1.00 (5)** / 0.00	0.26 (15) / 0.03	0.05 (15) / 0.03	**1.00 (6)** / 0.00
claude-haiku-4.5	1.00	0.20 (15) / 0.03	0.01 (15) / 0.04	**1.00 (5)** / 0.00	0.30 (15) / 0.04	0.04 (15) / 0.02	**1.00 (6)** / 0.00

too short for OOB's slow ONS-with-ε to accumulate wealth past $1/\alpha$. Empirical FPR remains at or below $\alpha = 0.05$ uniformly, in agreement with Proposition 1.

D.3 Detector score functions

Each offline detector maps a text span y to a scalar score $\phi(y)$. These scores are then used as inputs to the kernel-witness payoff in TRIBET. Thus, the detector determines the one-dimensional representation in which evidence is constructed, while TIDS controls false positives through rank calibration. We consider four representative zero-shot detector scores.

Fast-DetectGPT (FDGPT). Fast-DetectGPT [13] compares the log-likelihood of the test text under a proxy LLM against the expected log-likelihood under an analytic sampling model around the same proxy distribution. We use the analytical sampling-discrepancy score

$$\phi_{\text{FDGPT}}(y) = \frac{\ell_\theta(y) - \mu_\theta(y)}{\sigma_\theta(y)},$$

where $\ell_\theta(y) = \log p_\theta(y)$ and $\mu_\theta(y), \sigma_\theta(y)$ are the mean and standard deviation of the proxy log-likelihood under its next-token distribution. FDGPT is the strongest detector setting in Chen and Wang [21] and is our default score in Table 2.

DetectGPT (DetGPT). DetectGPT [12] is a perturbation-curvature detector. Given $K = 30$ T5-mask perturbations $\tilde{y}_1, \ldots, \tilde{y}_K$, it computes

$$\phi_{\text{DetGPT}}(y) = \log p_\theta(y) - \frac{1}{K}\sum_{k=1}^{K} \log p_\theta(\tilde{y}_k).$$

Machine-generated text often lies near a local maximum of the proxy likelihood, leading to a larger likelihood drop under perturbations. These scores can be heavy-tailed when the perturbation distribution and proxy distribution diverge on short or unusual spans.

NPR. NPR [14] replaces the DetectGPT likelihood drop with a perturbation-rank statistic:

$$\phi_{\text{NPR}}(y) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\{\log p_\theta(\tilde{y}_k) > \log p_\theta(y)\}.$$

This rank normalization removes the absolute likelihood scale and bounds the score in $[0, 1]$, but the statistic can still be unstable when the perturbation set is highly variable.

Binoculars (Bino). Binoculars [15] compares the test span's perplexity under an observer model with its cross-perplexity under a performer model. We use the canonical Falcon-7B/Falcon-7B-Instruct pair recommended by the authors and sign-align the resulting score so that larger values correspond to stronger MGT evidence.

For DetectGPT and NPR, we run the full proxy sweep and report the best-AUROC proxy for each source, matching the convention of Table 2. For Binoculars, the canonical observer/performer pair is fixed.

D.4 Score-function robustness: full results on alternative scores

We replicate the main Olympic experiment using three alternative detector scores: DetectGPT, NPR, and Binoculars. Tables 7–9 follow the same schema as Table 2. All cells use 200 Monte-Carlo iterations at $\alpha = 0.05$ and $n_{\text{cal}} = 20$.

Power-reporting convention. For detector scores with near-random held-out AUROC, power comparisons are not informative: the score itself contains little usable MGT signal, so failures of power reflect weak evidence rather than calibration. We therefore suppress power and stopping-time entries for cells with $\hat{C}(f) < 0.63$, while still reporting empirical FPR. This separates the two questions addressed by the tables: whether TRIBET preserves FPR control for arbitrary detector scores, and whether it converts informative scores into power when sufficient rank separation is present.

Table 7: DetectGPT score robustness on Olympic self-reference and XSum→Olympic cross-domain streams. Each cell follows the format of Table 2; cells with $\hat{C}(f) < 0.63$ report FPR only. TRIBET controls FPR across all cells, while power is reported only when the detector score has sufficient held-out separability.

		Self-reference (Olympic)			Cross-domain (XSum→Olympic)		
Source LLM	$\hat{C}(f)$	OOB-S1	OOB-S2	**TriBet**	OOB-S1	OOB-S2	**TriBet**
gpt4o-mini	0.83	1.00 (55) / 0.03	0.87 (130) / 0.01	**1.00 (8)** / 0.01	1.00 (64) / 0.02	1.00 (77) / **0.13**†	**1.00 (9)** / 0.01
gpt4o	0.78	1.00 (62) / 0.03	0.69 (146) / 0.01	**1.00 (10)** / 0.00	1.00 (65) / 0.02	1.00 (83) / **0.26**†	**1.00 (12)** / 0.01
gpt5.4	0.61	– (–) / 0.02	– (–) / 0.01	– (–) / 0.00	– (–) / 0.02	– (–) / **0.51**†	– (–) / 0.01
gpt5.4-pro	0.61	– (–) / 0.03	– (–) / 0.01	– (–) / 0.04	– (–) / 0.02	– (–) / **0.05**†	– (–) / 0.00
gemini-2.5-flash	0.73	1.00 (73) / 0.03	0.40 (160) / 0.01	**1.00 (17)** / 0.01	1.00 (91) / 0.01	1.00 (113) / **0.11**†	**1.00 (23)** / 0.00
gemini-2.5-pro	0.69	1.00 (118) / 0.02	0.08 (168) / 0.01	**1.00 (35)** / 0.02	1.00 (143) / 0.02	0.99 (193) / 0.02	**1.00 (42)** / 0.01
gemini-3.1-flash	0.71	1.00 (86) / 0.02	0.08 (169) / 0.01	**1.00 (26)** / 0.00	1.00 (97) / 0.02	1.00 (111) / **0.24**†	**1.00 (35)** / 0.01
gemini-3.1-pro	0.60	– (–) / 0.03	– (–) / 0.01	– (–) / 0.01	– (–) / 0.02	– (–) / **0.15**†	– (–) / 0.00
claude-opus-4.6	0.74	1.00 (77) / 0.02	0.26 (161) / 0.01	**1.00 (16)** / 0.02	1.00 (82) / 0.02	1.00 (111) / **0.08**†	**1.00 (20)** / 0.00
claude-sonnet-4.6	0.65	0.96 (137) / 0.02	0.03 (173) / 0.01	**1.00 (60)** / 0.00	1.00 (141) / 0.02	1.00 (195) / 0.05	**1.00 (101)** / 0.03
claude-opus-4.5	0.77	1.00 (77) / 0.04	0.76 (140) / 0.01	**1.00 (12)** / 0.01	1.00 (85) / 0.02	1.00 (94) / **0.16**†	**1.00 (14)** / 0.02
claude-haiku-4.5	0.73	1.00 (83) / 0.03	0.44 (157) / 0.01	**1.00 (18)** / 0.01	1.00 (93) / 0.02	1.00 (97) / **0.35**†	**1.00 (17)** / 0.01

Table 8: NPR score robustness on Olympic self-reference and XSum→Olympic cross-domain streams. Each cell follows the format of Table 2; cells with $\hat{C}(f) < 0.63$ report FPR only. TRIBET controls FPR across all cells, while power is reported only when the detector score has sufficient held-out separability.

		Self-reference (Olympic)			Cross-domain (XSum→Olympic)		
Source LLM	$\hat{C}(f)$	OOB-S1	OOB-S2	**TriBet**	OOB-S1	OOB-S2	**TriBet**
gpt4o-mini	0.81	1.00 (59) / 0.02	0.56 (148) / 0.02	**1.00 (10)** / 0.01	1.00 (60) / 0.02	1.00 (81) / **0.09**†	**1.00 (12)** / 0.00
gpt4o	0.76	1.00 (69) / 0.03	0.24 (166) / 0.02	**1.00 (13)** / 0.01	1.00 (75) / 0.02	1.00 (94) / **0.11**†	**1.00 (28)** / 0.01
gpt5.4	0.62	– (–) / 0.03	– (–) / 0.02	– (–) / 0.02	– (–) / 0.02	– (–) / **0.35**†	– (–) / 0.01
gpt5.4-pro	0.57	– (–) / 0.03	– (–) / 0.02	– (–) / 0.02	– (–) / 0.02	– (–) / 0.03	– (–) / 0.01
gemini-2.5-flash	0.71	1.00 (96) / 0.02	0.03 (173) / 0.02	**1.00 (31)** / 0.01	1.00 (106) / 0.02	1.00 (121) / **0.11**†	**1.00 (33)** / 0.01
gemini-2.5-pro	0.67	1.00 (116) / 0.03	0.03 (173) / 0.02	**1.00 (77)** / 0.01	1.00 (127) / 0.02	0.97 (142) / **0.07**†	**1.00 (82)** / 0.00
gemini-3.1-flash	0.70	1.00 (110) / 0.02	0.02 (173) / 0.02	**1.00 (40)** / 0.02	1.00 (129) / 0.02	1.00 (122) / **0.16**†	**1.00 (65)** / 0.00
gemini-3.1-pro	0.58	– (–) / 0.03	– (–) / 0.02	– (–) / 0.01	– (–) / 0.02	– (–) / **0.24**†	– (–) / 0.01
claude-opus-4.6	0.75	1.00 (76) / 0.03	0.14 (169) / 0.02	**1.00 (18)** / 0.04	1.00 (80) / 0.02	0.98 (139) / 0.02	**1.00 (32)** / 0.01
claude-sonnet-4.6	0.66	0.97 (143) / 0.03	0.02 (173) / 0.02	**1.00 (49)** / 0.04	1.00 (154) / 0.02	0.81 (290) / 0.02	**1.00 (93)** / 0.00
claude-opus-4.5	0.78	1.00 (71) / 0.03	0.56 (150) / 0.02	**1.00 (14)** / 0.02	1.00 (72) / 0.02	1.00 (96) / 0.03	**1.00 (34)** / 0.00
claude-haiku-4.5	0.71	1.00 (118) / 0.03	0.03 (173) / 0.02	**1.00 (46)** / 0.01	1.00 (156) / 0.01	0.03 (486) / 0.01	**1.00 (41)** / 0.00

Summary of Tables 7–9. The additional detector-score results support the same separation observed in the main experiments. FPR control remains stable across all three score families, including perturbation-based DetectGPT and NPR scores that are challenging for offset-calibrated betting. Power, however, depends on the detector's separability. When $\hat{C}(f)$ clears the reporting threshold, TRIBET typically achieves high power with shorter stopping time; when the score is near-random, no

Table 9: Binoculars score robustness on Olympic self-reference and XSum→Olympic cross-domain streams. Each cell follows the format of Table 2; cells with $\hat{C}(f) < 0.63$ report FPR only. $\hat{C}(f)$ is the held-out Olympic AUROC of the sign-aligned Binoculars score.

		Self-reference (Olympic)			Cross-domain (XSum→Olympic)		
Source LLM	$\hat{C}(f)$	OOB-S1	OOB-S2	**TriBet**	OOB-S1	OOB-S2	**TriBet**
gpt4o-mini	0.87	1.00 (50) / 0.02	0.93 (115) / 0.02	**1.00 (7)** / 0.00	1.00 (146) / 0.02	0.70 (304) / 0.02	**1.00 (10)** / 0.01
gpt4o	0.81	1.00 (61) / 0.02	0.65 (138) / 0.02	**1.00 (10)** / 0.02	0.76 (420) / 0.02	0.04 (456) / 0.02	**1.00 (17)** / 0.00
gpt5.4	0.64	0.85 (151) / 0.02	0.02 (172) / 0.02	0.94 (108) / 0.02	0.01 (498) / 0.02	0.01 (498) / 0.02	**1.00 (179)** / 0.00
gpt5.4-pro	0.56	–(–)/0.03	–(–)/0.02	–(–)/0.02	–(–)/0.02	–(–)/0.02	–(–)/0.01
gemini-2.5-flash	0.60	–(–)/0.02	–(–)/0.02	–(–)/0.01	–(–)/0.02	–(–)/0.02	–(–)/0.03
gemini-2.5-pro	0.74	1.00 (81) / 0.02	0.21 (156) / 0.02	**1.00 (20)** / 0.00	0.02 (498) / 0.02	0.01 (498) / 0.02	**1.00 (37)** / 0.03
gemini-3.1-flash	0.73	1.00 (90) / 0.03	0.11 (162) / 0.02	**1.00 (26)** / 0.01	0.01 (498) / 0.02	0.01 (498) / 0.02	**1.00 (29)** / 0.00
gemini-3.1-pro	0.50	–(–)/0.02	–(–)/0.02	–(–)/0.00	–(–)/0.02	–(–)/0.02	–(–)/0.01
claude-opus-4.6	0.83	1.00 (55) / 0.03	0.80 (127) / 0.02	**1.00 (9)** / 0.00	1.00 (271) / 0.02	0.16 (390) / 0.02	**1.00 (10)** / 0.01
claude-sonnet-4.6	0.72	1.00 (93) / 0.02	0.10 (164) / 0.02	**1.00 (28)** / 0.01	0.01 (498) / 0.02	0.01 (498) / 0.02	**1.00 (46)** / 0.01
claude-opus-4.5	0.80	1.00 (64) / 0.03	0.59 (142) / 0.02	**1.00 (10)** / 0.01	0.18 (470) / 0.02	0.03 (487) / 0.02	**1.00 (13)** / 0.01
claude-haiku-4.5	0.69	1.00 (111) / 0.02	0.05 (165) / 0.01	**1.00 (47)** / 0.01	0.01 (498) / 0.02	0.01 (498) / 0.02	**1.00 (46)** / 0.01

calibration method can manufacture reliable evidence. Thus, detector quality determines how much signal is available, while TIDS determines whether the resulting online test remains calibrated.

D.5 Mixed-MGT streams: varying machine-content fraction

We next evaluate partially contaminated streams, where only a fraction p of each document is machine-generated. This setting is closer to realistic misuse, where a document may contain a mixture of human-written and generated spans. We vary $p \in \{5, 10, 20, 30, 50, 70, 100\}\%$ on Olympic self-reference streams, using the same per-source proxy as in Table 2. The case $p = 100\%$ coincides with the fully machine-generated setting.

Table 10: Mixed-MGT streams under Olympic self-reference using FastDetectGPT and deployable TriBet AD. Each cell reports Pow/FPR over 300 Monte-Carlo trials at $\alpha = 0.05$ and $n_{\text{cal}} = 200$. Bold marks Pow ≥ 0.99. Lower p corresponds to stronger dilution by human-written text.

Source LLM	$p=5\%$	$p=10\%$	$p=20\%$	$p=30\%$	$p=50\%$	$p=70\%$	$p=100\%$
gpt4o-mini	0.00/0.00	0.00/0.00	0.84/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00
gpt4o	0.00/0.01	0.00/0.01	0.69/0.01	**1.00**/0.01	**1.00**/0.01	**1.00**/0.01	**1.00**/0.01
gpt5.4	0.00/0.00	0.00/0.00	0.02/0.00	0.01/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00
gemini-2.5-flash	0.00/0.00	0.00/0.00	0.00/0.00	0.04/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00
gemini-2.5-pro	0.00/0.01	0.00/0.01	0.04/0.01	0.99/0.01	**1.00**/0.01	**1.00**/0.01	**1.00**/0.01
gemini-3.1-flash	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.26/0.00	**1.00**/0.00	**1.00**/0.00
gemini-3.1-pro	0.01/0.00	0.00/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00
claude-opus-4.6	0.00/0.00	0.01/0.00	0.90/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00
claude-sonnet-4.6	0.00/0.00	0.03/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00
claude-opus-4.5	0.00/0.00	0.00/0.00	0.02/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00
claude-haiku-4.5	0.00/0.01	0.00/0.01	0.00/0.01	0.06/0.01	**1.00**/0.01	**1.00**/0.01	**1.00**/0.01

Analysis of Table 10. The main pattern is monotone and interpretable. FPR remains controlled across all contamination levels, confirming that null calibration is not weakened by the alternative mixture proportion. Power decreases as the machine-generated fraction is diluted, but the transition point depends on detector separability: high-AUROC sources are detected at moderate contamination levels, while weaker sources require a larger machine-generated fraction before the wealth process accumulates enough evidence. This supports the central interpretation of TriBet: validity is provided by conformal calibration, whereas power is governed by the separation induced by the detector score.

D.6 Sensitivity to the rejection level α

We evaluate sensitivity to the rejection level by sweeping $\alpha \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ on a representative subset of source LLMs (from strong-signal source LLM to weak-signal source LLM). This experiment is intended as a robustness check rather than a precise calibration study, since the number of trials is limited. All other parameters are held fixed.

Table 11: α-sweep on cross-domain FastDetectGPT streams. Each cell reports $\mathrm{Pow}/\mathrm{FPR}$ over 20 trials. TRIBET remains below or near the target level across the sweep, while OOB-S2 shows faster FPR growth on several cross-domain cells. Because the trial count is small, this table should be read as a sensitivity check rather than a precise estimate of tail calibration.

Source LLM	Method	$\alpha = 0.01$	0.02	0.05	0.1	0.2	0.5
palm2	OOB-S2	1.00/0.00	1.00/0.00	1.00/0.03	1.00/0.06	1.00/0.09	1.00/0.24
	TRIBET	1.00/0.00	1.00/0.00	1.00/0.01	1.00/0.01	1.00/0.06	1.00/0.12
gpt4o	OOB-S2	0.96/0.00	0.98/0.01	0.98/0.03	0.98/0.04	0.98/0.09	0.99/0.18
	TRIBET	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.02	1.00/0.10
gpt5.4	OOB-S2	0.01/0.00	0.01/0.01	0.01/0.04	0.04/0.04	0.04/0.09	0.11/0.18
	TRIBET	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.01	1.00/0.01	1.00/0.10
gpt5.4-pro	OOB-S2	0.01/0.01	0.01/0.01	0.01/0.03	0.02/0.04	0.04/0.09	0.07/0.20
	TRIBET	1.00/0.00	1.00/0.00	1.00/0.01	1.00/0.01	1.00/0.04	1.00/0.12
claude-sonnet-4.6	OOB-S2	0.01/0.00	0.01/0.01	0.02/0.03	0.06/0.04	0.09/0.09	0.21/0.17
	TRIBET	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.01	1.00/0.01	1.00/0.08
gemini-3.1-pro	OOB-S2	0.77/0.00	0.81/0.00	0.90/0.04	0.92/0.04	0.94/0.07	0.97/0.15
	TRIBET	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.01	1.00/0.01	1.00/0.09

E Future work.

TRIBET is a first step toward bridging the anytime-valid guarantees of game-theoretic statistics and the practical demands of online MGT auditing. Two natural directions remain open. First, the present analysis relies on assumptions that could be relaxed, including conditional exchangeability between the calibration set and the test stream, admissibility of the payoff, sufficient detectability of the score-induced ranks, and predictable choices of κ_t. Beyond-exchangeability conformal prediction [50] offers tools for quantifying coverage degradation under distribution shift, while recent results in game-theoretic statistics [23] suggest that predictable or exchangeability-preserving kernel selection and online bandwidth adaptation may be incorporated without sacrificing anytime validity. Second, the present AUROC$\to M$ analysis is intentionally coarse. Sharper kernel-aware sufficient conditions, and a more precise characterization of when the kernel witness induces large concordance and hence large conformal log-rank exponent M, would enlarge the regime in which TRIBET is provably effective. Beyond these theoretical questions, the betting framework can accommodate richer task-specific payoffs, including stream-aware payoffs that exploit document structure, multi-author payoffs for partial authorship at the span level, and watermark-aware payoffs when signed cryptographic side information is available. Each such variant would induce a new wealth process, and the same conformal e-process would preserve anytime validity as long as the payoff remains admissible and is applied symmetrically to calibration and test points. We hope this payoff–calibrator decoupling will serve as a general template for future online MGT auditing methods.

F Limitations and responsible use.

Exact anytime validity requires round-wise conditional super-uniformity; finite-pool reuse is supported empirically here. Calibration data must represent the monitored human population. Power depends on the detector and witness, degrades for low machine-content fractions, and is not evaluated here against real paraphrasing or humanization attacks. A wealth trajectory is evidence against a stream-level null, not a per-span authorship probability. Detection outputs should support human review rather than serve as sole evidence of misconduct.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the two contributions explicitly: (i) a triple-sample relative kernel-witness payoff, and (ii) a Test-Inclusive Domain-Shift conformal e-process with adaptive online κ_t. Both are precisely formalised in Sections 4 and Appendix B, and the empirical claims (anytime-α FPR control across 24 Olympic cells, 24 Essay cells, and 72 alternative-detector cells; faster stopping than OOB on every cell where both methods control FPR) are supported by Tables 2, 4, 6, and 7–9.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 and the appendix discuss three limitations: (i) the AUROC$\rightarrow M$ bridge is necessary but not sufficient for kernel-witness admissibility, as illustrated by the boundary cell discussed in §D.2; (ii) on cells with $\hat{C}(f) < 0.63$, no online method delivers reliable power, although TriBet retains FPR validity; (iii) the adaptive κ_t schedule controls rate, not anytime-α validity, so its empirical performance depends on the predictable estimator of $\hat{M}_t$.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The three propositions composing Theorem 1 are stated with explicit assumptions (Definitions 1–2 and Assumption 2); complete proofs appear in Appendix C, and each lemma cited within the proofs is itself proved or referenced to a published source.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 and Appendix D.1 document the data corpora, source LLMs, generation prompts, length-parity truncation rule, scoring procedure, and detector hyperparameters (n_{cal}, n_{pilot}, $T_{\max}$, α, $\kappa_{\text{lo/hi}}$). All hyperparameters are reported numerically, and the submitted supplemental material includes a self-contained `reproducible_code/` folder that reproduces Table 2 end-to-end given the publicly available Fast-DetectGPT scores, switching to the other score function mentioned in the Appendix can generate other results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A `reproducible_code/` folder is provided in the supplemental material. The folder contains all detection drivers required to reproduce Table 2, an environment specification (`requirements.txt`), and a top-level `README.md` documenting how to download the proxy LLM checkpoints from Hugging Face and how to obtain the XSum, Olympic-2024, and Student Essay corpora. Source-LLM and proxy-LLM model identifiers are listed verbatim in Appendix D.1.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 reports all detection-time hyperparameters ($\alpha = 0.05$, $n_{\text{cal}} = 20$, kernel bandwidth via the median heuristic, κ initialised at 0.5 and clipped to $[0.05, 0.995]$). Appendix D.1 documents the generation-time choices (temperature 1.0, `max_new_tokens` $=$ 300 for news and 1500 for essays, 30-token prefix for the news-prefix protocol). Train/test splits are not applicable because TriBet is a non-parametric streaming detector with no learned parameters; the calibration and reference splits are described in Section 4.2.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All cells in Tables 2, 6, and 7–9 report empirical Power, $\bar{\tau}$, and FPR computed over 200 Monte-Carlo iterations per cell (300 for the mixed-MGT table, with 3 cal-split seeds $\times$ 100 iterations). FPR violations are flagged with the marker † for cells whose empirical FPR exceeds $\alpha = 0.05$ at significance $P < 0.05$ under a one-sided binomial test on 200 trials. For two boundary cells the appendix reports a 10-cal-seed marginal estimate (marker ‡) to control single-seed variance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Detector scoring is performed on a single AMD MI325X GPU (256 GB HBM3E) per proxy LLM; one 500-article scoring run takes approximately 30 seconds. The detection step (kernel witness, conformal e-process, betting) runs on CPU and takes between 0.5 and 3 minutes per cell audit, depending on the stream length. The total compute footprint of the full evaluation is approximately 80 GPU-hours and 400 CPU-hours; preliminary and discarded experiments add roughly the same amount, primarily on the proxy-sweep search.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research uses publicly available human-text corpora (BBC XSum, an Olympic-news collection from major outlets, and the MGTBench Student Essay subset), all licensed for academic use, and proprietary or open-weight LLMs accessed under their respective terms of service. No personally identifiable information is collected or released. The work conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The introduction motivates the work by the deployment risks of misclassifying human-written documents as machine-generated (the Deloitte audit case at 20-span and 200-span document scales) and by the public-health risks of undetected machine-generated misinformation. The proposed framework provides anytime-valid FPR control, which directly mitigates the false-accusation risk; we discuss potential dual-use concerns and the residual dependence of detection power on proxy choice in Section 5.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper does not release any new pre-trained model or scraped dataset that poses a high misuse risk. The proxy LLMs used at scoring time (Falcon, Gemma, Llama, GPT-Neo) are released under their respective vendor licences, and the source-LLM machine-text outputs are stored only as token-level detector scores in the supplemental material; raw machine text is not redistributed.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets, codebases, and pre-trained model weights are cited at first use. The XSum corpus is released under the MIT licence by Narayan et al. [28]; the Olympic-2024 articles are public news under fair-use academic terms; the Student Essay corpus is released by He et al. [46] under the MGTBench academic licence. Open-weight proxy LLMs are accessed under their vendor licences (Llama Community Licence, Falcon TII Licence, Gemma Terms of Use, Apache-2.0 for GPT-Neo). The Online-Detection codebase of Chen and Wang [21] is built upon under its MIT licence.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: The supplemental material contains the TRIBET reference implementation under `reproducible_code/`, accompanied by a top-level `README.md` that describes the data layout, environment requirements, and exact commands to reproduce Table 2. The Olympic-2024 machine-text scores are documented in Appendix D.1 together with the generation prompts.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects. All human-text corpora used in the evaluation are public-domain articles or pre-existing essay collections released by their original authors.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects and therefore does not require IRB approval.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs play two non-standard roles in this work: (i) twelve frontier source LLMs (OpenAI, Google, Anthropic) are used as the *generators* of the machine-text side of every detection corpus; (ii) eight open-weight LLMs (Llama, Falcon, Gemma, GPT-Neo) are used as *proxy* models in the Fast-DetectGPT, DetectGPT, NPR, and Binoculars score functions. All API identifiers, prompts, and proxy choices are documented in Appendix D.1. LLMs were not used to develop the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.