RESOLVING SCORE-ORDERING REVERSAL IN WATERMARKED TEXT: AN ASYMMETRIC TWO-SIDED TEST FOR LLM TEXT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM watermarking and LLM-generated text detection are two complementary techniques for misinformation detection, intellectual property protection, and related provenance tasks. However, we found the two conflict: the watermarked LLM text can reverse the score-ordering of the statistical score by the state-of-the-art logits-based detectors, leading to failure in LLM text detection. We dive into this phenomenon to propose a new statistical measure for LLM-generated text and design a two-sided test accordingly. The test is enhanced threefold: 1) we design the upper and lower test branches separately with a pair of asymmetric statistical measures; 2) we remove the tail of the reference distribution due to the observation that both human and LLM rarely output low-probability tokens; 3) we use the context entropy to weigh each token position differently in the statistical measure for better discrimination. On a total of 660 configuration settings (LLMs $\times$ watermarking methods $\times$ datasets), our detector exceeds Fast-DetectGPT and AdaDetectGPT in terms of mean AUROC by 0.1929 and 0.1986, showing higher detection rates on a wider range of LLM text.

1 INTRODUCTION

LLM watermarking and LLM text detection are two complementary techniques for provenance tracing: the former proactively embeds identifiable signals into the text during generation by shifting the output distribution (Kirchenbauer et al., 2023; Zhao et al., 2024; Lee et al., 2024), while the latter evaluates a piece of observed text against the reference distribution to determine if it is an LLM generation (Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024). The state-of-the-art LLM text detectors mostly leverage the observation that text sampled from an LLM tends to occupy negative curvature regions of the model's log probability function (Mitchell et al., 2023) and thus employ curvature-based criteria for detection (Mitchell et al., 2023; Bao et al., 2024; Miao et al., 2024; Zhou et al., 2025).

We found that such curvature-based criteria are no longer reliable on watermarked text: LLM watermarking can induce a **score-ordering reversal** in the detector evidence, causing machine-generated text to score below human writing. In other words, watermarked LLM text can appear more human-like than real human writing in terms of the detector score. To make things worse, text watermarked by different LLM watermarking methods do not share a uniform evidence according to the current logit-based, log-probability detectors. Figure 1 illustrates this score-ordering reversal with OPT-2.7B serving as the scoring model. Unwatermarked text, EXP, and SynthID (Pan et al., 2024; Dathathri et al., 2024) score above human writing, whereas CGW and HeavyWater (Christ et al., 2024; Tsur et al., 2025) score below it. Hence we ask a question: Can we design a robust statistical test for LLM-generated text, regardless of whether it is watermarked?

To answer this question, a natural starting point is a two-sided test based on the detector score. Such a test can capture machine text that falls on either side of the human score distribution. Yet the opposite score shifts observed above indicate that watermarking does not always affect the two sides symmetrically. Enforcing both tails to refer to the same distribution would miss all the useful, one-sided evidences. Hence we instead construct the upper- and lower-tail rejection regions separately.

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

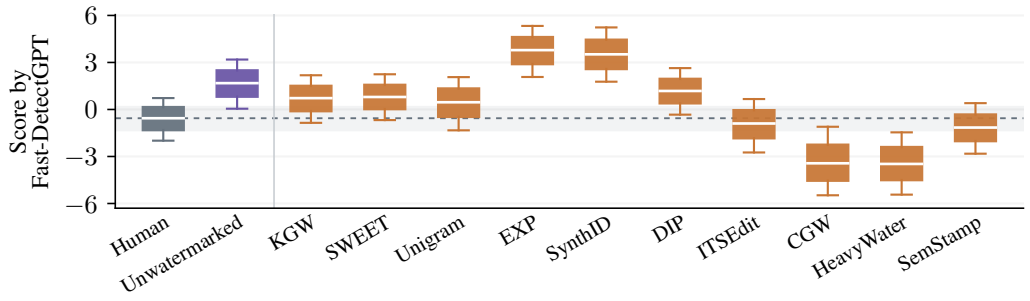


Figure 1: Score distribution for LLM text detection by FastDetectGPT on different watermarking methods. OPT-2.7B serves as the scoring model. Boxes/whiskers: 25th–75th/10th–90th percentiles. Gray band/dashes: human interquartile range/median.

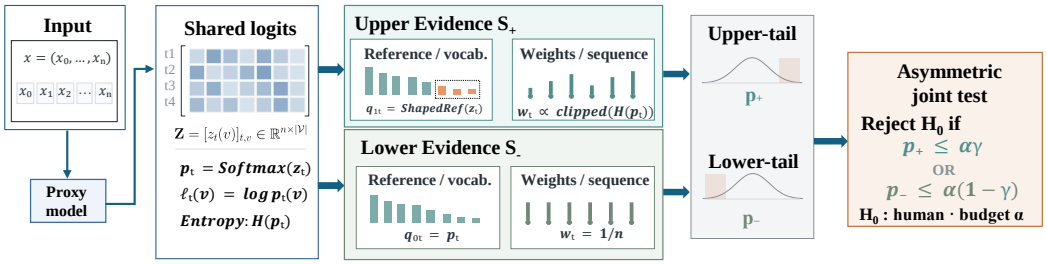


Figure 2: Our asymmetric two-sided test framework. A proxy model supplies token likelihoods to both the upper- and lower- tail test. The lower-tail test statistic is the same with FastDetectGPT while the upper-tail one prioritizes high-entropy positions and excludes low-probability tokens from the reference distribution for better discrimination. Human tail ranks and a fixed error-budget split combine the two statistics for the test.

It is our observation that watermarking has selective impact to the token distribution. SWEET (Lee et al., 2024), for example, applies its watermark bias only at high-entropy positions, indicating non-uniform watermark signal distribution. More generally, when the context strongly constrains the next token, such a token choice would hardly distinguish human or machine generation. In contrast, when several continuations remain plausible, differences in their token preferences can be a strong distinction signal. Furthermore, low-probability candidates offer limited discriminative power, as they are consistently absent in both human- and machine-generated texts. Therefore, we prioritize higher-entropy positions while suppressing low-probability token candidates to shape the reference distribution for the upper-tail rejection region. Specifically, we exclude both highly certain positions and highly unlikely candidate tokens to prevent skewing the reference distribution. This shaped distribution is not applied to the lower-tail rejection region, as emphasizing high-entropy positions would otherwise compromise the validity of the lower-tail inference.

Based on these observations, we propose an asymmetric two-sided test. Figure 2 summarizes our proposed detector. Given a passage, a proxy model produces the token probabilities used in both test branches in a single forward pass. To construct the upper-tail test statistic, we truncate the token distribution at cumulative mass c, and assign context-entropy weights to different positions to calculate the statistics. The lower-tail test statistic reuses the statistic measure adopted in FastDetectGPT. On the targeted text, we compute two scores for both the upper- and lower-tail test statistics, yielding two tail p-values which are combined in a two-sided test.

Across **660 settings** spanning five generators, twelve datasets, ten watermarking methods, and unwatermarked generation, our detector achieves a mean AUROC of **0.8219** with a single OPT-2.7B proxy model serving as both the sampling model and the scoring model, and **0.8317** with Falcon-7B as the sampling model and Falcon-7B-Instruct as the scoring model. Our method improves the mean AUROC over FastDetectGPT and AdaDetectGPT by 0.1929 and 0.1986, respectively (Table 1). Against a modified version of Fast-DetectGPT with a symmetric two-sided test by their

test statistic, the gains remain around 0.05 (Table 4), showing the effectiveness of our asymmetric statistics construction.

Our contributions are:

- **Identifying score-ordering reversal phenomenon for watermarked text.** We document this phenomenon across watermarked LLM pipelines, and analyze how different watermarking methods, decoding settings, and generator-proxy differences contribute to the observed score distributions.
- **Designing an asymmetric two-sided test.** We formulate the upper- and lower-tail test statistics separately. The upper-tail statistic is computed using entropy-weighted positions and a truncated distribution to enhance the discriminative power between human- and machine-generated texts.
- **Evaluation across a diverse range of LLMs and watermarking methods** show our detector exceeds the state-of-the-art detectors by a large margin, verifying the effectiveness of our detector on both watermarked and unwatermarked text.

2 Preliminaries and Related Work

Let the tokenized document in detection be $\mathbf{x} = (x_0, \ldots, x_n)$, where each token is from vocabulary $\mathcal{V}$. By feeding the prefix $\mathbf{x}_{<t}$ into the proxy LLM, we obtain the logit vector $\mathbf{z}_t = (z_t(v))_{v\in\mathcal{V}}$ and the corresponding next-token distribution $p_t(v) = \mathrm{Softmax}(\mathbf{z}_t)_v$. We denote the token log likelihood by $\ell_t(v) = \log p_t(v)$. Fixing the observed prefix $\mathbf{x}_{<t}$, we also draw a token y_t from a sampling model following distribution q_t. The Fast-DetectGPT statistic (Bao et al., 2024) is

$$S_{\mathrm{FD}}(\mathbf{x}) = \frac{\sum_{t=1}^{n}[\ell_t(x_t) - \mathbb{E}_{q_t}\ell_t(y_t)]}{\sqrt{\sum_{t=1}^{n}\mathrm{Var}_{q_t}(\ell_t(y_t))}}. \tag{1}$$

Its prior work DetectGPT (Mitchell et al., 2023) does not sample from q_t but rather from p_t by feeding the proxy LLM with perturbed passages $\tilde{\mathbf{x}}_{<t}$ to estimate the expectation and variance in Eq. (1). AdaDetectGPT (Zhou et al., 2025) learns a witness function $w(l_t)$ to replace l_t in Eq. (1) to optimize the distinguishing power. Other statistics compare likelihood with log rank (Su et al., 2023), normalize observed log-perplexity by two-model cross-entropy (Hans et al., 2024), or measure normalized temperature sensitivity (Ma et al., 2026).

The distribution of l_t depends on the underlying text generation process. Watermarking can alter this distribution by changing how tokens or sentences are selected, potentially affecting the detectors described above. To see where these changes arise, consider a generation step: the model computes logits from current context, converts them into softmax probabilities by which the next token is sampled. KGW intervenes at the logit stage with a context-dependent greenlist (Kirchenbauer et al., 2023); SWEET embeds watermarking signals at high-entropy positions (Lee et al., 2024), while Unigram uses a fixed vocabulary partition (Zhao et al., 2024) for watermarking. Dipmark and HeavyWater instead operate on token probabilities. Dipmark reweighs the distribution while preserving its output expectation over watermark randomness (Wu et al., 2024), whereas HeavyWater uses heavy-tailed scores and optimal transport to construct watermark-conditioned distributions (Tsur et al., 2025). Other methods intervene during sampling: CGW uses keyed pseudorandomness (Christ et al., 2024), ITSEdit couples inverse-transform sampling with a watermark key sequence (Kuditipudi et al., 2024), and SynthID selects tokens through keyed tournaments (Dathathri et al., 2024). SemStamp operates on complete candidate sentences through semantic hashing and rejection sampling (Hou et al., 2024). Appendix A discusses how each watermarking mechanism may affect the detectors considered in this work.

3 Method

As aforementioned, watermarked text may cause score-ordering reversal in previous statistical measures and change their detection decisions. Observing the asymmetric score distribution around the human writings, it is clear that a single-sided test would miss a part of the evidence. Using

one statistic for both tails imposes the same reference distribution and position weights, which may strengthen evidence in one tail while weakening it in the other.

In our observation, not all token positions distinguish human and machine text equally well. When the context strongly favors the same continuation for both, their token choices provide little distinction. Where several alternatives remain plausible, differences in their preferences may be more revealing. This motivates us to emphasize higher-entropy positions in the statistical measure. Moreover, we found the low-probability token hardly contributes to the distinction as both human and machine-generated text rarely output them. Removing them can be helpful in focusing on the more discriminative part of the distribution. Hence, our approach imposes a tail-truncated distribution for computing the statistic, while giving higher priorities to positions in a more uncertain context.

Let $q_{0t} = p_t$ denote the original proxy distribution and q_{1t} be the shaped distribution derived from q_{0t}. We define the moments of ℓ_t as

$$\mu_{jt} = \mathbb{E}_{q_{jt}} \ell_t, \qquad \sigma_{jt}^2 = \mathrm{Var}_{q_{jt}}(\ell_t), \qquad j \in \{0, 1\}. \tag{2}$$

Thus, $(\mu_{0t}, \sigma_{0t}^2)$ are the moments under the original reference, while $(\mu_{1t}, \sigma_{1t}^2)$ are those under the shaped reference. With nonnegative position weights w_t, the token residuals $\ell_t(x_t) - \mu_{jt}$ give

$$S_j = \frac{\sum_t w_t[\ell_t(x_t) - \mu_{jt}]}{\max\{\sqrt{\sum_t (w_t)^2 \sigma_{jt}^2}, \varepsilon\}}. \tag{3}$$

The reference determines the mean and variance used for comparison, while the position weights determine how strongly each token contributes to the document-level response.

Distribution shaping by nucleus sampling. We construct the upper-tail reference by applying temperature scaling followed by top-c truncation (top-p in (Holtzman et al., 2020)):

$$q_{1t} = \mathrm{Softmax}_c\Big(\frac{\mathbf{z}_t}{T}\Big). \tag{4}$$

Here $\mathrm{Softmax}_c$ denotes the softmax operation followed by truncation at cumulative mass c: tokens are retained in decreasing probability order until the cumulative mass first exceeds c, including the crossing token, and the retained probabilities are renormalized. We use $T = 1.2$ and $c = 0.8$ by default. The shaped reference is used only to compute the upper-tail statistics. Each observed token retains its original proxy log probability $\ell_t(x_t)$, even if it is excluded by the top-c truncation.

Position weights are based on the context entropy calculated using the output token distribution as follows:

$$H(p_t) = -\frac{\sum_{v \in \mathcal{V}} p_t(v)\ell_t(v)}{\log|\mathcal{V}|}, \qquad \bar{H} = \frac{1}{n}\sum_{t=1}^{n} H(p_t). \tag{5}$$

We normalize the entropy at each position that $H(p_t) \in [0, 1]$ as $\log|\mathcal{V}|$ is the maximum entropy of a distribution over $\mathcal{V}$. We cap the per-position entropy at κ times the mean entropy: values below the cap remain unchanged and larger values are clipped to the cap. Each position weight is the proportion of the entropy in the entire context:

$$w_t = \frac{\min\big\{H(p_t), \max\big(\kappa\bar{H}, \varepsilon_h\big)\big\}}{\max\big\{\sum_{i=1}^{n} \min\big\{H(p_i), \max\big(\kappa\bar{H}, \varepsilon_h\big)\big\}, \varepsilon_w\big\}}, \tag{6}$$

Here $\kappa = 1.5$, $\varepsilon_h = 10^{-6}$, and $\varepsilon_w = 10^{-8}$ by default.

Upper-tail test statistics. Combining the shaped reference in Eq. (4) with the position weights in Eq. (6), we define the upper-tail test statistics as

$$S_+(\mathbf{x}) = \frac{\sum_{t=1}^{n} w_t\,[\ell_t(x_t) - \mu_{1t}]}{\max\Big\{\sqrt{\sum_{t=1}^{n} w_t^2 \sigma_{1t}^2}, \varepsilon\Big\}}. \tag{7}$$

A larger S_+ indicates a stronger upper-tail evidence, with the observed tokens deviating upward from the shaped reference after position weighting and variance normalization. This test branch targets machine text that scores above human writing. Since watermarking can also move machine text below the human score distribution, we construct the lower-tail test statistics separately.

Lower-tail test statistics. We observe that in the lower-tail rejection region, low-entropy positions can still contain observed tokens which contributes to the distinguishing power. Thus we keep $q_{0t} = p_t$ and $w_t = 1/n$, giving

$$S_-(\mathbf{x}) = \frac{n^{-1}\sum_t [\ell_t(x_t) - \mu_{0t}]}{\max\{n^{-1}\sqrt{\sum_t \sigma_{0t}^2}, \varepsilon\}}. \tag{8}$$

A smaller S_- indicates a stronger lower-tail evidence: the observed tokens receive lower proxy log probabilities than expected under the original reference, relative to its reference variability. Because all positions contribute equally, low-entropy positions are not downweighted, preserving negative residuals that may be suppressed by the entropy weighting used in the upper branch. Section 5.1 shows that this choice retains useful separation for CGW and HeavyWater. With the numerical floor inactive, S_- reduces to the Fast-DetectGPT statistic in Equation (1) with proxy model and sampling model to be the same.

Human tail calibration. We calibrate the two test scores against human documents. Let $\mathbf{x}_1^{\mathrm{H}}, \ldots, \mathbf{x}_m^{\mathrm{H}}$ be $m \geq 1$ calibration documents from unique human sources, and store $\mathcal{C} = \{(S_+(\mathbf{x}_i^{\mathrm{H}}), S_-(\mathbf{x}_i^{\mathrm{H}}))\}_{i=1}^m$. The inclusive, add-one empirical tail p-values are (Angelopoulos & Bates, 2023)

$$p_+(\mathbf{x}) = \frac{1 + \sum_{i=1}^m \mathbf{1}\{S_+(\mathbf{x}_i^{\mathrm{H}}) \geq S_+(\mathbf{x})\}}{m+1}, \qquad p_-(\mathbf{x}) = \frac{1 + \sum_{i=1}^m \mathbf{1}\{S_-(\mathbf{x}_i^{\mathrm{H}}) \leq S_-(\mathbf{x})\}}{m+1}. \tag{9}$$

A small p_+ means that few calibration human texts have an equal or higher upper-tail test score, while a small p_- means that few have an equal or lower lower-tail test score. We use $m = 800$ and recompute the human calibration scores whenever a branch's reference or weighting rule changes.

Two-sided test decision rule. We aggregate the calibrated tail p-values using a total nominal false-positive budget allocated for human text $\alpha \in (0, 1)$. The upper-tail test receives $\alpha\gamma$ and the lower-tail test $\alpha(1-\gamma)$, where $\gamma \in (0, 1)$. A document is classified as machine-generated when either test meets its rejection threshold:

$$\varphi_\alpha(\mathbf{x}) = \mathbf{1}\{p_+(\mathbf{x}) \leq \alpha\gamma \quad \text{or} \quad p_-(\mathbf{x}) \leq \alpha(1-\gamma)\}. \tag{10}$$

We use $\gamma = 1/2$, assigning equal budgets to the two branches, and evaluate both branches for every document.

To sum up, we provide the detail in Algorithm 1 which implements the proxy model to be same with the sampling model. Appendix C gives the default settings, numerical details and extension to the Falcon pair.

4 Experimental Results

4.1 Experimental Setup

Calibration data. We construct the calibration set from C4 (Raffel et al., 2020), SQuAD (Rajpurkar et al., 2016), Writing (Fan et al., 2018), and XSum (Narayan et al., 2018), paired with five generators to form 20 dataset-generator combinations. Each combination includes unwatermarked generation and ten watermarking methods, giving 220 dataset–generator–condition settings for parameter selection. We select T, c, κ by AUROC on this set and use its 800 unique human documents, 200 per dataset, to calibrate the two score tails. All detector configurations are fixed before test evaluation.

Test data. The test set contains 2,360 unique human documents from 12 datasets and 129,800 machine-generated texts from the same five generators. Each dataset–generator combination includes the same eleven generation conditions, yielding $12 \times 5 \times 11 = 660$ evaluation settings. The calibration and test sets contain disjoint human sources. The test datasets cover news, dialogue, discussions, reviews, science, narratives and commonsense. Table 9 lists their sources and sample counts; Appendix C.3 provides generator details and the data separation procedure.

Algorithm 1 An asymmetric two-sided test for machine-generated detection

Input: Tokenized document $\mathbf{x}$ with $n \geq 1$ prediction positions; proxy model M; human score bank $\mathcal{C} = \{(S_+(\mathbf{x}_i^{\mathrm{H}}), S_-(\mathbf{x}_i^{\mathrm{H}}))\}_{i=1}^m$; budgets $\alpha, \gamma \in (0,1)$; fixed parameters $T, c, \kappa, \varepsilon_h, \varepsilon_w, \varepsilon$.
Output: machine-text decision $\varphi_\alpha(\mathbf{x})$.

$(\mathbf{z}_t)_{t=1}^n \leftarrow M(\mathbf{x})$ ▷ One forward pass
for $t = 1, \ldots, n$ **do**
  $p_t(v) \leftarrow \mathrm{Softmax}(\mathbf{z}_t)_v$, $\ell_t(v) \leftarrow \log p_t$ for $v \in \mathcal{V}$
  $q_{0t} \leftarrow p_t$; $q_{1t} \leftarrow \mathrm{Softmax}_c(\mathbf{z}_t/T)$ ▷ Original and shaped references
  for $j \in \{0,1\}$ **do** ▷ Reference moments of ℓ_t
    $\mu_{jt} \leftarrow \sum_{v\in\mathcal{V}} q_{jt}(v)\ell_t(v)$
    $\sigma_{jt}^2 \leftarrow \sum_{v\in\mathcal{V}} q_{jt}(v)[\ell_t(v) - \mu_{jt}]^2$
  end for
  $H(p_t) \leftarrow -\sum_{v\in\mathcal{V}} p_t(v)\ell_t(v)/\log|\mathcal{V}|$ ▷ Normalized proxy entropy
end for
Compute $\bar{H}$ by Eq. (5) ▷ Mean entropy across positions
Compute w_t by Eq. (6) ▷ Cap and normalize entropy
Compute $S_+(\mathbf{x})$ by Eq. (7) ▷ Use the shaped reference and entropy weights
Compute $S_-(\mathbf{x})$ by Eq. (8) ▷ Use the original reference and uniform weights
Compute $p_+(\mathbf{x})$ and $p_-(\mathbf{x})$ by Eq. (9) ▷ Empirical upper- and lower-tail p-values
Compute $\varphi_\alpha(\mathbf{x})$ by Eq. (10) ▷ Reject on either tail
return $\varphi_\alpha(\mathbf{x})$

Detector configurations. Fast-DetectGPT (FD; Bao et al., 2024), AdaDetectGPT (Ada; Zhou et al., 2025), and our Falcon variant use Falcon-7B (Almazrouei et al., 2023) as the sampling model and Falcon-7B-Instruct as the scoring model. Binoculars (Hans et al., 2024) uses the same pair as its observer and performer, respectively. The configurations of Ada and Binoculars (†) are detailed in Appendix C.3. NTS (Ma et al., 2026) and the likelihood, entropy, rank and log-rank baselines use a single Falcon-7B proxy model; NTS applies temperatures of 0.7 and 1.4. Our single-proxy configuration uses OPT-2.7B (Zhang et al., 2022) as both the sampling model and the scoring model. Input limits are 512 tokens for NTS and 2,048 for the other detectors. We report each baseline's original score for the main comparison. Two-sided FD and Ada controls share the proxy models, human calibration documents and $\gamma = 1/2$ split with the corresponding variants of our detector.

Metrics. For AUROC and ROC evaluation, we rank documents for our two-sided detector using

$$J(\mathbf{x}) = -\log \min\left\{\frac{p_+(\mathbf{x})}{\gamma}, \frac{p_-(\mathbf{x})}{1-\gamma}\right\}. \tag{11}$$

Larger values indicate stronger evidence of machine generation. We report AUROC averaged across the 660 evaluation settings, the mean AUROC over each detector's worst-performing 66 settings (CVaR_{10}), and the number of settings with reversed score ordering (SOR; AUROC $< 1/2$). TPR@1%FPR and TPR@5%FPR are the highest observed TPRs at test-human FPRs no greater

Table 1: Performance over 660 evaluation settings. Bold: column best, including ties (SOR minimized); red: AUROC or CVaR_{10} below 0.5. TPR is expressed in percentages. †: configuration details in Appendix C.3.

Detector	AUROC	CVaR_{10}	SOR	TPR@1%FPR	TPR@5%FPR
Likelihood	0.5295	0.0180	293	16.81	25.52
Entropy	0.5449	0.0468	283	10.57	20.06
Rank	0.5702	0.0801	242	13.28	24.30
Log-Rank	0.5452	0.0210	285	17.87	26.64
Fast-DetectGPT	0.6388	0.0179	228	29.94	38.90
AdaDetectGPT†	0.6331	0.0195	220	28.30	36.33
Binoculars†	0.6135	0.0263	249	22.68	32.27
NTS	0.7341	0.3041	107	29.22	39.08
Ours (OPT)	0.8219	0.5473	13	**39.35**	**52.56**
Ours (Falcon pair)	**0.8317**	**0.5835**	**0**	31.40	52.29

Table 2: Mean AUROC by generator (each result is averaged over 12 dataset $\times$ 11 generation conditions settings). Bold, red and † follow Table 1.

Detector	GLM-4-9B(0414)	Falcon-7B	Llama-2-7B	OPT-2.7B	OPT-6.7B
Likelihood	0.6052	0.6179	0.5701	0.4115	0.4426
Entropy	0.4321	0.5063	0.5381	0.6380	0.6100
Rank	0.5657	0.6371	0.5647	0.5308	0.5524
Log-Rank	0.6121	0.6394	0.5782	0.4327	0.4635
Fast-DetectGPT	0.6453	0.6781	0.6590	0.6009	0.6106
AdaDetectGPT†	0.6500	0.6710	0.6589	0.5887	0.5970
Binoculars†	0.6122	0.6467	0.6428	0.5771	0.5889
NTS	0.7033	0.6993	0.8275	0.7263	0.7138
Ours (OPT)	0.7838	0.7878	0.8469	**0.8631**	**0.8279**
Ours (Falcon pair)	**0.8169**	**0.8061**	**0.8627**	0.8469	0.8258

Table 3: Mean AUROC for different watermarking methods (each result is averaged over 12 datasets under unwatermarked generation and ten watermarking methods). In computing AUROC, Upper uses $-\log p_+$ and Two-sided uses J. Bold: row best, including ties; red and † follow Table 1.

	Baselines			Ours (OPT)		Ours (Falcon pair)	
Mechanism	FD	Ada†	NTS	Upper	Two-sided	Upper	Two-sided
Unwatermarked	0.9141	0.8990	0.8239	**0.9331**	0.8803	0.9295	0.9053
KGW	0.7861	0.7462	0.6499	**0.8584**	0.7880	0.8566	0.8204
SWEET	0.8067	0.7682	0.6849	0.8602	0.7782	**0.8604**	0.8215
Unigram	0.7023	0.6563	0.5798	**0.7940**	0.7450	0.7867	0.7532
EXP	0.9970	**0.9974**	0.9702	0.9934	0.9882	0.9956	0.9942
SynthID	0.9918	**0.9924**	0.9673	0.9894	0.9802	0.9894	0.9857
DIP	0.8612	0.8358	0.7475	**0.8942**	0.8180	0.8858	0.8488
ITSEdit	0.4204	0.4603	0.5218	0.5005	0.6217	0.5403	**0.6274**
CGW	0.1084	0.1250	0.7933	0.2024	**0.9077**	0.2633	0.8907
HeavyWater	0.0812	0.0887	0.7887	0.1465	**0.9114**	0.1978	0.8790
SemStamp	0.3573	0.3950	0.5472	0.4322	**0.6222**	0.4771	0.6221

than 1% and 5%, respectively, averaged across different configuration settings. We also report mean power using thresholds set by the human calibration bank, together with the observed FPR, counting each unique test human document once. Paired intervals are obtained by resampling whole datasets (Appendix C.3).

4.2 Detection Results

Overall detection performance. Our asymmetric two-sided test reduced the score-ordering reversals that undermine fixed-direction detection (Table 1). Across 660 evaluation settings, SOR fell from 107 for NTS to 13 for OPT and 0 for Falcon. Both variants also achieved worst-decile mean AUROCs above chance (0.5473/0.5835 versus 0.3041), showing improved average separation where each detector performed worst. Overall mean AUROC reached 0.8219/0.8317 for OPT/Falcon, versus NTS's 0.7341.

Generator and dataset results. Both proxy configurations outperformed every baseline across all five generators (Table 2). Compared with NTS, AUROC gains ranged from 0.0194 to 0.1368 for OPT and from 0.0352 to 0.1206 for Falcon. OPT and Falcon also exceeded the strongest baseline on seven and ten of twelve datasets, respectively. Entropy led on CNN/DailyMail and NTS on WritingPrompts (Appendix D, Table 10).

Upper and lower branches across watermarks. The lower branch recovered separation under score reversal, complementing the upper branch's gains on unwatermarked text (Table 3). The upper branch achieved unwatermarked AUROCs of 0.9331/0.9295 for OPT/Falcon, exceeding FD (0.9141) and Ada (0.8990). Adding the lower branch raised OPT AUROC from 0.2024 to 0.9077 on

CGW and from 0.1465 to 0.9114 on HeavyWater. Both exceeded NTS (0.7933/0.7887), while FD and Ada remained below 0.13. Using both tails reduced unwatermarked AUROC to 0.8803/0.9053 for OPT/Falcon. ITSEdit and SemStamp remained difficult, with AUROC near 0.62, while Ada led on EXP and SynthID.

Table 4: Two-sided detection over 660 evaluation settings with 800 calibration humans and $\gamma = 1/2$. FD/Ada use a symmetric score for both tails. Power/FPR are percentages at nominal 1%/5% budgets; FPR counts unique test humans. Bold: best within each proxy (AUROC/power maximized, FPR minimized); † follows Table 1.

Configuration	AUROC	Power_1	FPR_1	Power_5	FPR_5
FD + two tails (OPT)	0.7809	37.06	**1.27**	52.63	7.50
Ours (OPT)	**0.8219**	**40.32**	1.31	**54.86**	**6.69**
FD + two tails (Falcon)	0.7768	40.26	2.58	49.90	**6.23**
Ada + two tails (Falcon)†	0.7572	39.66	3.05	47.30	6.57
Ours (Falcon pair)	**0.8317**	**44.54**	**2.29**	**55.83**	6.53

Comparison with two-sided baselines. This comparison holds the lower-tail test statistics, proxy, calibration sources and error allocation fixed, changing only the upper-tail test statistics to be the symmetric of the lower one. Our asymmetric test improved AUROC over two-sided FD by 0.0410 with OPT and 0.0549 with Falcon (Table 4), demonstrating gains beyond two-sided symmetric test. At nominal 1%, power/FPR (%) was 40.32/1.31 for OPT and 44.54/2.29 for Falcon, versus FD's 37.06/1.27 and 40.26/2.58, respectively. Paired 95% confidence intervals for both AUROC gains excluded zero (Table 11).

5 Component Analysis

We first changed the upper-branch weights, then its reference distribution, while keeping the lower statistic, 800 human calibration sources and $\gamma = 1/2$ fixed. We recomputed upper calibration scores and empirical tail p-values for each configuration (Table 5).

Table 5: Staged upper-branch ablation with the lower statistic and calibration sources fixed. Original/uniform denotes two-sided FD. Power and observed FPR are percentages at a nominal total level of 1%, with $\gamma = 1/2$. Bold indicates the highest AUROC for each proxy.

		OPT			Falcon pair		
Upper reference	Weights	AUROC	Power_1	FPR_1	AUROC	Power_1	FPR_1
Original	Uniform	0.7809	37.06	1.27	0.7768	40.26	2.58
Original	Entropy	0.8191	40.87	1.23	0.8134	40.89	1.86
Shaped	Entropy	**0.8219**	40.32	1.31	**0.8317**	44.54	2.29

5.1 Position Weighting

With the original reference, replacing uniform upper-branch weights with entropy weights increased mean joint AUROC from 0.7809 to 0.8191 for OPT. Falcon's AUROC increased from 0.7768 to 0.8134 under the same change. Entropy weighting produced the largest gain in this staged ablation of the complete detector.

Under the shaped reference, entropy weighting increased the unstandardized machine–human mean residual difference in every unwatermarked evaluation setting. For OPT, the mean difference rose from 0.3854 to 0.5251 nats. With OPT's original reference, entropy weighting reduced the human–machine mean residual differences by 0.1141 nats for CGW and 0.0766 nats for HeavyWater (Table 6). These measurements support different position weights for the two branches (Appendix E.1). Proposition 2 characterizes shared optimal weights for the standardized mean effects under fixed contexts. The directional mean shifts, divided by their respective reference variances, must be proportional across positions. Appendix E.3 reports detection sensitivity to the entropy cap.

Table 6: Lower-branch detection and ablations. AUROC uses $-S_-$, with the original reference and uniform weights. Δ_{U-E} measures the reduction in the human–machine mean residual difference under entropy weighting (nats). $\mathcal{L}_{\text{center}}$ measures the loss in pairwise ranking credit when only the reference center changes, assigning half-credit to ties. Positive losses favor the original lower-branch construction.

Proxy	Mechanism	Lower AUROC	Δ_{U-E}	$\mathcal{L}_{\text{center}}$
OPT	CGW	0.9125	0.1141	0.0397
OPT	HeavyWater	0.9372	0.0766	0.0359
Falcon	CGW	0.8916	0.0803	0.0692
Falcon	HeavyWater	0.9188	0.0417	0.0609

5.2 Reference Shaping

With entropy weights fixed, shaping the upper reference increased joint AUROC by 0.0028 for OPT and 0.0183 for Falcon (Table 5). The paired 95% intervals were $[-0.0037, 0.0123]$ and $[0.0059, 0.0320]$, respectively, providing clearer evidence of a gain for Falcon. Worst-decile mean AUROC also increased for both proxies (Table 15). At nominal 1%, Falcon's power and FPR both increased, while OPT's power decreased slightly. OPT's AUROC improved on CGW/HeavyWater and ITSEdit/SemStamp but declined on unwatermarked text (Table 14). Changing T or c by 0.1 reduced OPT AUROC to 0.7429–0.7837, showing sensitivity to the joint parameter choice (Appendix E.4.2).

For the lower branch, changing only the reference center at the original scale reduced OPT's pairwise ranking credit by 0.0397 on CGW and 0.0359 on HeavyWater (Table 6). A larger correction to human than machine scores narrows the lower-tail gap (Equation (33)). These results support the asymmetric reference design: shaping the upper reference while retaining the original reference for the lower branch.

5.3 Branch Contributions and Calibration

With equal allocation at nominal 1%, exclusive upper- and lower-branch detections contributed 30.40 and 9.93 percentage points to OPT's macro-averaged power. Falcon's corresponding contributions were 36.40 and 8.14 percentage points under the same allocation. For OPT, exclusive upper-branch detection reached 48.22% on unwatermarked text, compared with 0.05% for the lower branch. Exclusive lower-branch detection reached 49.62% on CGW and 50.60% on HeavyWater (Table 17). Adding the lower branch preserves all upper rejections at a fixed upper threshold. It can still change the joint ranking, as the unwatermarked AUROC comparison shows (Section 4.2).

Parameter selection and calibration shared development human texts; Appendix C.3 details the assumptions of Proposition 5. At nominal 1%, OPT produced 0 false positives among 581 human texts with 1–64 prediction positions, and 7 among 258 with more than 512. These length groups contain different domain mixtures; Appendix F reports the full length and calibration-bank analyses.

6 Conclusion

We address score-ordering reversal with an asymmetric two-sided test. The upper branch uses reference shaping and entropy weights; the lower branch uses the original reference and uniform weights. In the single-model implementation, both responses share one OPT forward pass and are combined using human tail ranks. Across 660 tasks, the test improves mean AUROC over matched two-sided Fast-DetectGPT by 0.0410 with OPT and 0.0549 with Falcon. These results support choosing references and position weights separately when machine text can rank on either side of human writing.

AI USE STATEMENT

Generative AI assisted with organizing and interpreting experimental results, explaining detector score behavior, and formulating mathematical statements and proofs. It also supported implementation audits and the preparation of manuscript text, experimental scripts, plotting code, LaTeX and references. AI-assisted checks compared numerical claims with recorded outputs, verified source and cache hashes, and checked references against primary sources. We ran the experimental code, checked the raw outputs and references, and reviewed the mathematical statements, proofs and final manuscript. Tables and experimental figures were derived from recorded model outputs. Our analyses used cached scores, new OPT forward passes for parameter sensitivity, and 360 controlled HeavyWater generations using prompts from calibration datasets. Machine-generated benchmark texts came from the generator and watermark implementations documented in Appendix C.3.

ETHICS STATEMENT

Machine-generated text detectors can incorrectly label human writing as machine-generated, potentially exposing writers to unfair accusations. This risk is particularly relevant when detectors are applied beyond their calibration domains. We report human false-positive rates across domains and text lengths, together with detection performance under different generation conditions. Detector scores should be considered alongside other evidence when assessing authorship. Consequential applications require calibration for the intended domain, human review and a process for contesting errors. The study used existing human texts and machine-generated continuations and did not recruit human participants.

REPRODUCIBILITY STATEMENT

All detector hyperparameters were selected on the calibration set and fixed for test evaluation. Section 3 and Algorithm 1 specify the detector. Appendices A–B.3 provide the mechanism and construction arguments; Appendix C collects fixed settings, numerical conventions and calibration. Appendix C.3 records the evaluation protocol, and Appendix D extends the detection results. Appendices E–F follow the three questions in Section 5: directional weighting, reference shaping, and reliability across domains and lengths.

REFERENCES

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The Falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023. doi: 10.48550/arXiv.2311.16867.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations*, pp. 24814–24836, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6b8c6f846c3575e1d1ad496abea28826-Abstract-Conference.html.

Hong Chen, Hiroya Takamura, and Hideki Nakayama. SciXGen: A scientific paper dataset for context-aware text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1483–1492, 2021a. doi: 10.18653/v1/2021.findings-emnlp.128.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 5062–5074, 2021b. doi: 10.18653/v1/2021.findings-acl.449.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1125–1139. PMLR, 2024. URL https://proceedings.mlr.press/v247/christ24a.html.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. Scalable watermarking for identifying large language model outputs. *Nature*, 634: 818–823, 2024. doi: 10.1038/s41586-024-08025-4.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, 2018. doi: 10.18653/v1/P18-1082.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3558–3567, 2019. doi: 10.18653/v1/P19-1346.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. In *International Conference on Machine Learning*, pp. 17519–17537, 2024. URL https://proceedings.mlr.press/v235/hans24a.html.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. SemStamp: A semantic watermark with paraphrastic robustness for text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4067–4082, 2024. doi: 10.18653/v1/2024.naacl-long.226.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019. doi: 10.18653/v1/D19-1259.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084, 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=FpaCL1MO2C.

Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. Who wrote this code? watermarking for code generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4890–4911, 2024. doi: 10.18653/v1/2024.acl-long.268.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36–53, 2024. doi: 10.18653/v1/2024.acl-long.3.

Shixuan Ma, Jiahao Li, Zhendong Mao, and Quan Wang. Zero-shot detection of LLM-generated text using temperature sensitivity. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37664–37679, 2026. doi: 10.18653/v1/2026.acl-long.1748.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, 2011. URL https://aclanthology.org/P11-1015/.

Yibo Miao, Hongcheng Gao, Hao Zhang, and Zhijie Deng. Efficient detection of LLM-generated texts with a Bayesian surrogate model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6118–6130, 2024. doi: 10.18653/v1/2024.findings-acl.366.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pp. 24950–24962, 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, 2016. doi: 10.18653/v1/N16-1098.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290, 2016. doi: 10.18653/v1/K16-1028.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, 2018. doi: 10.18653/v1/D18-1206.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. MarkLLM: An open-source toolkit for LLM watermarking. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 61–71, 2024. doi: 10.18653/v1/2024.emnlp-demo.7.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016. doi: 10.18653/v1/D16-1264.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12395–12412, 2023. doi: 10.18653/v1/2023.findings-emnlp.827.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 613–624, 2016. doi: 10.1145/2872427.2883081.

Team GLM, Aohan Zeng, Bin Xu, et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. URL https://arxiv.org/abs/2406.12793.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Dor Tsur, Carol Long, Claudio Mayrink Verdun, Sajani Vithana, Hsiang Hsu, Richard Chen, Haim Permuter, and Flavio Calmon. HeavyWater and SimplexWater: Distortion-free LLM watermarks for low-entropy distributions. In *Advances in Neural Information Processing Systems*, volume 38, pp. 107204–107256. Curran Associates, Inc., 2025. doi: 10.52202/085713-3575.

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. A resilient and accessible distribution-preserving watermark for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53443–53470, 2024. URL https://proceedings.mlr.press/v235/wu24h.html.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019. doi: 10.18653/v1/P19-1472.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. doi: 10.48550/arXiv.2205.01068.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. In *International Conference on Learning Representations*, pp. 43738–43772, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/beae9ed5316bcc48e616754c06c11875-Abstract-Conference.html.

Hongyi Zhou, Jin Zhu, Pingfan Su, Kai Ye, Ying Yang, Shakeel Gavioli-Akilagun, and Chengchun Shi. AdaDetectGPT: Adaptive detection of LLM-generated text with statistical guarantees. In *Advances in Neural Information Processing Systems*, volume 38, pp. 89077–89118, 2025. doi: 10.52202/085713-2980.

A Watermark Generation and Score-Ordering Reversal

This appendix explains Figure 1 through the generation mechanisms in Section 2. We distinguish each watermark operation from its decoding inputs and the proxy used for detection.

A.1 Generation Operations and Local Score Changes

We distinguish logit bias, probability reweighting, keyed sampling, transport coupling and sentence acceptance. Each operation acts on an input distribution that may already include decoding adjustments.

At a fixed history h, let G_h and $W_{\omega,h}$ be the input and output laws of watermark operation ω. The history includes the prompt and relevant state. For a finite score probe ℓ_h, the local change is

$$\Delta_\omega(h) = \langle W_{\omega,h} - G_h, \ell_h \rangle. \tag{12}$$

With matched vocabularies and contexts, ℓ_h is the token log probability. Different tokenizers require a probe on the appended string or the document-level analysis in Appendix A.6.

A local shift identity. For the fixed-prefix analysis in Appendix A, a finite token preference ψ defines

$$G_\eta(v) = \frac{G_h(v)e^{\eta\psi(v)}}{\sum_u G_h(u)e^{\eta\psi(u)}}. \tag{13}$$

Proposition 1 (Local score shift)**.** *For a finite probe f and fixed G_h, ψ, f independent of η,*

$$\frac{d}{d\eta}\mathbb{E}_{G_\eta} f = \mathrm{Cov}_{G_\eta}(f, \psi). \tag{14}$$

Proof. Differentiating the normalizer gives $\partial_\eta G_\eta(v) = G_\eta(v)[\psi(v) - \mathbb{E}_{G_\eta}\psi]$. Multiplying by $f(v)$ and summing proves the identity. □

For $f = \ell_h$, the covariance determines the local direction. Appendix A.6 analyzes document-level AUROC under the full generation process.

A.2 Favored-Token Bias

KGW and Unigram boost context-dependent and fixed greenlists, respectively; SWEET applies its boost above a generator-entropy threshold (Kirchenbauer et al., 2023; Zhao et al., 2024; Lee et al., 2024). For greenlist $\mathcal{G}$, mass $\pi = G_h(\mathcal{G})$ and boost $\eta > 0$, splitting the normalizer gives

$$\Delta_{\mathrm{green}}(h) = \frac{(e^\eta - 1)\pi(1-\pi)}{1 + (e^\eta - 1)\pi} \left[\mathbb{E}_{G_h}(\ell_h \mid \mathcal{G}) - \mathbb{E}_{G_h}(\ell_h \mid \mathcal{G}^c)\right]. \tag{15}$$

For $0 < \pi < 1$, the shift is negative exactly when greenlist tokens have lower mean proxy scores than the complement. Degenerate lists give zero change; a closed SWEET gate leaves the distribution unchanged.

A.3 Probability Reweighting and Keyed Sampling

An operation preserves its input on average if $\mathbb{E}_Z W_{\omega,h,Z} = G_h$ for fresh state Z independent of that input. This gives zero average local shift. Sequence preservation additionally requires this identity at every reachable history, with matched prompts and stopping rules. Reused keys can couple state and history.

DIP. DiPmark redistributes mass along a keyed vocabulary permutation σ (Wu et al., 2024). For token v's cumulative-probability interval I_v^σ and threshold α_{D}, the ideal output is

$$W_{\sigma,h}(v) = \int_{I_v^\sigma} \left[\mathbf{1}\{u > \alpha_{\mathrm{D}}\} + \mathbf{1}\{u > 1 - \alpha_{\mathrm{D}}\}\right] du. \tag{16}$$

Reversing the permutation gives $W_{\sigma,h} + W_{\mathrm{rev}(\sigma),h} = 2G_h$, so equiprobable reverse pairs preserve the input average. A fixed permutation or subsequent truncation can change proxy scores. Implementation settings are collected in Appendix C.3.

SynthID. One binary-score tournament maps input π to

$$\pi'(v) = \pi(v)\,[1 + g(v) - \mathbb{E}_\pi g]\,, \tag{17}$$

where $g(v)$ is the keyed score (Dathathri et al., 2024). Fresh scores independent of π, with equal token-wise means, give $\mathbb{E}_g \pi' = \pi$. The evaluated adapter reuses a finite score table and applies internal temperature scaling before outer truncation.

EXP. The evaluated MarkLLM loop uses the exponential race

$$v^* = \arg\max_v U_v^{1/G_h(v)} = \arg\min_v \frac{-\log U_v}{G_h(v)}. \tag{18}$$

Independent continuous uniform U_v make these exponential rates $G_h(v)$, giving $\Pr(v^* = v) = G_h(v)$ (Kuditipudi et al., 2024). Hashing four recent token IDs can reuse randomness at repeated or colliding contexts; retries also vary the generation temperature (Pan et al., 2024).

ITSEdit. Inverse-transform sampling selects the interval containing a uniform scalar in a permuted cumulative distribution (Kuditipudi et al., 2024). Fresh independent scalars preserve the input law, even with a fixed permutation. The evaluated loop uses a finite stream, clips end-of-sequence probabilities and bypasses standard sampling processors.

CGW. The MarkMyWords adapter selects token-code bits using their conditional probability masses (Christ et al., 2024). Fresh independent uniforms recover each token's input probability by multiplying its bit probabilities. The adapter instead keys decisions by the previous token and bit index, returning a one-hot token choice. It omits the original construction's initial random block and entropy-based block updates, so that construction's adaptive-query guarantee does not directly apply.

HeavyWater. HeavyWater selects a state-conditional optimal-transport (OT) column (Tsur et al., 2025). The evaluated path has no post-OT tilt; Appendix A.8 separates transport from decoding and state effects.

A.4 Sentence Acceptance

SemStamp accepts sentence proposals $Y \sim G_h$ according to a semantic-hash event A_h (Hou et al., 2024). With acceptance probability $a_{\mathrm{acc}} = G_h(A_h) > 0$, unlimited independent retries give $G_h(\cdot \mid A_h)$. For an integrable sentence score f_h,

$$\mathbb{E}_{G_h(\cdot|A_h)} f_h - \mathbb{E}_{G_h} f_h = \frac{\mathrm{Cov}_{G_h}(f_h, \mathbf{1}_{A_h})}{a_{\mathrm{acc}}}. \tag{19}$$

Acceptance lowers the mean exactly when the accepted proposals have lower mean scores. The evaluated wrapper uses three-bit hashes, two valid signatures, zero margin and at most $M_{\mathrm{retry}} = 100$ proposals. After failed retries, it retains the last proposal. For iid proposals, no empty-candidate stop and $0 < a_{\mathrm{acc}} < 1$, this fallback scales the shift by $1 - (1 - a_{\mathrm{acc}})^{M_{\mathrm{retry}}-1}$.

A.5 Decoding and Proxy Context

Preservation identities concern the distribution presented to the watermark, which can already include temperature scaling, truncation and repetition constraints. For generation temperature T_{gen}, distinct from reference temperature T, write $G_h^{[T_{\mathrm{gen}}]} \propto G_h^{1/T_{\mathrm{gen}}}$. Equation (14) gives

$$\frac{d}{dT_{\mathrm{gen}}} \mathbb{E}_{G_h^{[T_{\mathrm{gen}}]}} \ell_h = -\frac{\mathrm{Cov}_{G_h^{[T_{\mathrm{gen}}]}}(\ell_h, \log G_h)}{T_{\mathrm{gen}}^2}. \tag{20}$$

Thus lowering temperature raises the local mean when the covariance is positive. Top-k truncation also raises mean log probability with matched generation and scoring views; a different proxy can change this direction.

A repetition penalty $\rho > 1$ replaces a seen token's generator logit ζ_v by ζ_v/ρ if nonnegative, and by $\rho\zeta_v$ otherwise. Unseen logits are unchanged. The covariance identity in Equation (14) describes the resulting mean shift; reducing high-scoring tokens' logits can lower the proxy mean. HeavyWater applies this penalty before transport. Prompt omission, tokenizer differences and quality filtering further separate the generation and detection views.

A.6 From Generation Laws to Document AUROC

Local shifts affect later contexts and the final score normalization. To include these effects, define the human mid-distribution function for any score S:

$$F^{\mathrm{mid}}_{\mathrm{H},S}(s) = P_{\mathrm{H}}(S(\mathbf{X}) < s) + \tfrac{1}{2}P_{\mathrm{H}}(S(\mathbf{X}) = s). \tag{21}$$

Its machine expectation equals AUROC for higher-is-machine ordering, with half-credit for ties (Hanley & McNeil, 1982).

Fix a common prompt law and finite generation horizon, padding stopped sequences with an absorbing state. Let u_h and $W_{\omega,h}$ be baseline and watermarked next-token laws after marginalizing internal states. Define $F^{\mathrm{rank}}_{U,t}(h, v)$ as the expected final $F^{\mathrm{mid}}_{\mathrm{H},S_-}(S_-)$ after appending v and continuing under the baseline. This continuation is defined also at prefixes not reached by baseline generation. Then

$$\mathcal{A}_\omega - \mathcal{A}_U = \sum_t \mathbb{E}_{H_t\sim\omega}\langle W_{\omega,H_t} - u_{H_t}, F^{\mathrm{rank}}_{U,t}(H_t, \cdot)\rangle, \tag{22}$$

where $\mathcal{A}_U$ and $\mathcal{A}_\omega$ are the corresponding AUROCs. The identity follows by telescoping baseline continuation values along the watermarked trajectory. Different stopping rules are included in the kernels; different prompt laws require an additional initial-value term.

For $\mathcal{A}_U > 1/2$, reversal requires the right-hand side to be below $1/2 - \mathcal{A}_U$. Equal conditional laws preserve AUROC under matched prompts and stopping rules.

A.7 Observed Score Directions

Table 7 quantifies Figure 1 using macro AUROC across dataset–generator settings. Human scores were negative in 70.4% of documents.

Table 7: OPT score ordering in Figure 1. AUROC ranks by S_-, with larger values indicating machine text; the lower-tail test rejects small S_-. SOR counts reversed settings out of 60. Medians pool 2,360 humans or 11,800 texts per machine condition.

Mechanism	Median S_-	Macro AUROC	SOR
Human	-0.559	–	–
Unwatermarked	1.680	0.9155	0/60
KGW	0.723	0.7870	2/60
SWEET	0.796	0.8102	1/60
Unigram	0.460	0.7236	9/60
EXP	3.787	**0.9922**	0/60
SynthID	3.518	0.9902	0/60
DIP	1.181	0.8688	1/60
ITSEdit	-0.895	0.4169	41/60
CGW	-3.434	0.0875	60/60
HeavyWater	-3.476	0.0628	60/60
SemStamp	-1.154	0.3594	54/60

CGW and HeavyWater reversed ordering in all 60 settings, versus 41 for ITSEdit and 54 for SemStamp. EXP and SynthID remained above AUROC $1/2$ throughout. KGW, SWEET and Unigram shifted below unwatermarked text but retained higher overall medians than humans. ITSEdit's matched-OPT subgroup also retained mean AUROC above $1/2$.

A.8 HeavyWater: Preservation and Generation Controls

The adapter uses `tilt=False`, internal nucleus mass 0.999 and a fixed key indexed by the previous token (Tsur et al., 2025). Let a_h be its input after internal truncation. An exact OT plan P^*_h with

row marginal a_h and K equal column marginals gives conditional laws $q^*_{hs} = KP^*_h(:, s)$:

$$\frac{1}{K}\sum_{s=1}^{K} q^*_{hs} = a_h, \qquad \mathbb{E}_{Z\sim\mathrm{Unif}[K]}\mathbb{E}_{q^*_{hZ}}\ell_h = \mathbb{E}_{a_h}\ell_h. \tag{23}$$

Under fresh independent uniform states, the transport operation preserves its processed input a_h.

For numerical columns $\widehat{q}_{hs}$ with positive normalizers, put $\bar{q}_h = K^{-1}\sum_s \widehat{q}_{hs}$. Let u_h be the unwatermarked law, c_h the input before internal truncation, and $g(h)$ the keyed state. The local difference decomposes as

$$\begin{aligned}\langle \widehat{q}_{h,g(h)} - u_h, \ell_h\rangle = \langle c_h - u_h, \ell_h\rangle + \langle a_h - c_h, \ell_h\rangle \\ + \langle \bar{q}_h - a_h, \ell_h\rangle + \langle \widehat{q}_{h,g(h)} - \bar{q}_h, \ell_h\rangle.\end{aligned} \tag{24}$$

The terms separate decoding, truncation, numerical error and state selection. Averaging uniformly over states at a fixed prefix cancels the last term.

Generation controls. Recorded settings include repetition penalty 1.1 and no-repeat 4-grams. HeavyWater disables outer top-k; the unwatermarked call omitted it, invoking the environment default of 50. This default was reconstructed because some original configurations were not fully saved.

We generated 360 OPT-2.7B continuations under nine conditions on 40 calibration-partition prompts, ten per dataset. Source-ID hashes selected prompts independently of detector scores. Conditions shared prompt-specific seeds, batch size one, a 200-token maximum and a 64-token minimum, except the no-minimum control. Fresh states used a separate stream. All first outputs were retained, including eight quality-check failures. Each dataset AUROC compared ten machines with ten humans after common normalization; Table 8 averages these four AUROCs. The processed-input control retains internal truncation and its singleton convention without OT or state selection.

Table 8: Controlled generations on 40 calibration-partition prompts, scored by OPT S_- in the same higher-is-machine direction as Table 7. Except the two top-$k = 50$ rows, outer top-k is disabled. Repetition constraints mean penalty $\rho = 1.1$ and no-repeat 4-grams.

Generation condition	Mean S_-	AUROC
No watermark, top-$k = 50$	1.9640	**0.9575**
No watermark, unrestricted	-0.8155	0.3700
No watermark, repetition penalty 1.1	-3.4696	0.0275
No watermark, repetition constraints	-3.4521	0.0350
Processed input, without OT	-3.4709	0.0350
HeavyWater, fixed states	-3.6952	0.0125
HeavyWater, independent states	-3.6615	0.0050
HeavyWater, no repetition constraints	-0.2044	0.5525
No watermark, top-$k = 50$, no minimum	1.8289	0.9375

Removing top-k changed AUROC by -0.5875 (95% CI $[-0.6925, -0.4624]$); adding repetition penalty 1.1 changed it by -0.3425 ($[-0.4925, -0.2125]$). Intervals use 5,000 paired source-bootstrap resamples within datasets, conditional on one seed per source.

Reversal already occurred without OT. Adding fixed-state transport changed AUROC by -0.0225 ($[-0.0625, 0.0175]$), whereas adding repetition constraints to HeavyWater changed it by -0.5400 ($[-0.6700, -0.4150]$). The 4-gram and internal nucleus additions had small effects along these controls. Decoding therefore contributed to reversal independently of the watermark operation.

Context, length and acceptance checks. Among 11,559 first-attempt benchmark acceptances, mean OPT/Falcon AUROCs were 0.0626/0.0811, with reversal in all 60 settings. Prompt-conditioned scoring gave AUROCs 0.0550/0.0425 for processed-input/fixed-state outputs. Scoring the first 64 positions gave 0.1058/0.0775 using 40 machine texts and 39 eligible humans. Reversal thus persisted under matched context, common scored length and first-attempt acceptance.

B THEORY OF THE ASYMMETRIC TEST

We analyze the standardization and weighting choices in Sections 3 and 5.1, using the moments in Equation (2).

B.1 CONDITIONAL REFERENCE STANDARDIZATION

Fix the document prefixes, log probabilities, references and weights. For independent auxiliary draws $y_t \sim q_{jt}$, define

$$\sigma_j^2(w) = \sum_t w_t^2 \sigma_{jt}^2, \qquad s_j(w) = \max\{\sqrt{\sigma_j^2(w)}, \varepsilon\}, \qquad U_j = \sum_t w_t[\ell_t(y_t) - \mu_{jt}].$$

Independence gives $\mathbb{E}U_j = 0$ and $\mathrm{Var}(U_j) = \sigma_j^2(w)$. Thus $Z_j = U_j/s_j(w)$ satisfies

$$\mathbb{E}Z_j = 0, \qquad \mathrm{Var}(Z_j) = \frac{\sigma_j^2(w)}{\max\{\sigma_j^2(w), \varepsilon^2\}} \le 1. \tag{25}$$

These are auxiliary reference moments, including at zero variance; human score distributions are estimated from the calibration bank.

B.2 WHY THE TWO BRANCHES CAN FAVOR DIFFERENT WEIGHTS

Assume fixed contexts, positive reference variances, nonzero nonnegative weights and inactive numerical floors. The results concern standardized mean effects; AUROC also depends on human scores and generated contexts.

Let Q_t be a next-token law, with Q_t^+ and Q_t^- denoting two theoretical generation cases. Detection evaluates both branches without estimating these laws or assigning a direction. Define

$$\xi_t = \frac{\mathbb{E}_{Q_t^+}\ell_t - \mu_{1t}}{\sigma_{1t}^2}, \qquad \delta_t = \frac{\mu_{0t} - \mathbb{E}_{Q_t^-}\ell_t}{\sigma_{0t}^2}. \tag{26}$$

Positive values indicate departures above and below the respective reference means. By linearity, the mean upper and negative lower scores are

$$\mathcal{A}_+(w) = \frac{\sum_t w_t \xi_t \sigma_{1t}^2}{\sqrt{\sum_t w_t^2 \sigma_{1t}^2}}, \qquad \mathcal{A}_-(w) = \frac{\sum_t w_t \delta_t \sigma_{0t}^2}{\sqrt{\sum_t w_t^2 \sigma_{0t}^2}}. \tag{27}$$

Proposition 2 (Optimal and shared position weights)**.** *For a nonzero nonnegative vector $a = (a_t)$,*

$$\frac{\sum_t w_t a_t \sigma_{jt}^2}{\sqrt{\sum_t w_t^2 \sigma_{jt}^2}} \le \sqrt{\sum_t a_t^2 \sigma_{jt}^2}, \qquad \textit{with equality exactly when } w \propto a, \tag{28}$$

with positive proportionality. Hence nonnegative, nonzero ξ, δ admit a shared maximizing weight vector exactly when they are proportional.

Proof. Apply Cauchy–Schwarz to $(w_t\sigma_{jt})_t$ and $(a_t\sigma_{jt})_t$. Equality requires $w \propto a$ because all reference variances are positive. Taking $(j, a) = (1, \xi)$ and $(0, \delta)$ gives the two optima and their shared-weight condition. □

Directional cases. A constant positive δ_t makes uniform lower weights optimal. If $\xi_t = \lambda_+ w_t$ for $\lambda_+ > 0$ and entropy weights from Equation (6), those weights maximize the upper mean effect. When both conditions hold and entropy weights are nonconstant, sharing either choice strictly reduces the other mean effect.

When entropy weighting improves the upper mean. Let $\pi_{1t} = \sigma_{1t}^2/\sum_s \sigma_{1s}^2$, with w the entropy weights. For nonnegative, nonzero ξ, the upper mean exceeds that under uniform weights exactly when

$$\mathrm{Cov}_{\pi_1}(w, \xi) > \mathbb{E}_{\pi_1}\xi \left[\sqrt{\mathbb{E}_{\pi_1} w^2} - \mathbb{E}_{\pi_1} w\right]. \tag{29}$$

To see this, write $B_1 = \sum_t \sigma_{1t}^2$. The two effects are $\sqrt{B_1}\mathbb{E}_{\pi_1}(w\xi)/\sqrt{\mathbb{E}_{\pi_1} w^2}$ and $\sqrt{B_1}\mathbb{E}_{\pi_1}\xi$; expanding the covariance gives the condition. Section 5.1 tests the weighting choices through mean residual differences and complete-test AUROC.

B.3 Reference Shaping and Calibrated Decisions

We analyze the reference comparisons in Section 5.2 using Equation (3).

B.3.1 Reference Center and Scale

Write the shaped moments as $\mu_{1t}(T,c), \sigma^2_{1t}(T,c)$ and the distribution before truncation as $q_t^T = \mathrm{Softmax}(\mathbf{z}_t/T)$.

Proposition 3 (Reference-center monotonicity)**.** *For the single-model construction, $\mu_{1t}(T,c)$ is nonincreasing in $T > 0$ and separately in $c \in (0,1]$. On every interval where the nucleus support is fixed,*

$$\frac{\partial \mu_{1t}(T,c)}{\partial T} = -\frac{\sigma^2_{1t}(T,c)}{T^2} \le 0. \tag{30}$$

Proof. On fixed support, logits and ℓ_t differ by a constant, so differentiation gives Equation (30). Order tokens by decreasing ℓ_t, breaking ties by vocabulary index. For the first $K < |\mathcal{V}|$ tokens, their mass $C_K(T)$ satisfies

$$C_K'(T) = -T^{-2}\,\mathrm{Cov}_{q_t^T}(\mathbf{1}\{v \text{ is among the first } K\}, \ell_t) \le 0, \tag{31}$$

because their scores are no smaller than the rest. Raising T or c can only expand the nucleus prefix, adding lower-scoring tokens and therefore never increasing the center. □

Raising T and reducing c have opposing effects on the center; scores also depend on the reference variance. For the Falcon pair, distinct scoring and reference models replace the variance in this derivative with a cross-model covariance.

Human–machine gaps and score changes. For fixed weights, use $s_j(w)$ from Appendix B.1 and define

$$r = \frac{\sum_t w_t(\mu_{1t} - \mu_{0t})}{s_0(w)}, \qquad k = \frac{s_1(w)}{s_0(w)} > 0, \qquad S_1 = \frac{S_0 - r}{k}. \tag{32}$$

This identity includes the numerical floors and allows document-specific r, k. Holding the original scale fixed, let $U = S_0$ and $C = U - r$. Then

$$C_{\mathrm{H}} - C_{\mathrm{M}} = (U_{\mathrm{H}} - U_{\mathrm{M}}) - (r_{\mathrm{H}} - r_{\mathrm{M}}). \tag{33}$$

A larger human correction narrows the lower-tail margin. Setting $w_t = 1/n$ gives the center-only comparison in Section 5.2; rescaling by k can further change ordering.

For the upper comparison, fix entropy weights, lower scores, allocation and calibration sources; write $d_{\mathrm{sh}} = S_1 - S_0$. A common r and positive k across calibration and query documents would preserve all empirical tail p-values, including ties.

B.3.2 Movement Relative to the Human Rejection Boundary

Let $p_{+,j}$ be the upper-tail p-value from Equation (9), using S_j for the query and the same calibration humans. Write $P_{\mathrm{H}}, P_{\mathrm{M}}$ for the human and machine document laws conditional on this bank.

Proposition 4 (Calibrated shaping boundary)**.** *Fix the scoring rules, calibration documents, α and γ. For positive k_+ from Appendix C.2, let τ_j be the k_+th largest calibration upper score under reference j. Define*

$$m_0(\mathbf{x}) = S_0(\mathbf{x}) - \tau_0, \qquad e_\alpha(\mathbf{x}) = d_{\mathrm{sh}}(\mathbf{x}) - (\tau_1 - \tau_0). \tag{34}$$

The original and shaped upper rules reject exactly when $m_0(\mathbf{x}) > 0$ and $m_0(\mathbf{x}) + e_\alpha(\mathbf{x}) > 0$, respectively. With the lower rejection event $L_\alpha = \{\mathbf{x} : p_-(\mathbf{x}) \le \alpha(1-\gamma)\}$ unchanged, the complete test gains and loses detections on

$$\begin{aligned} \mathcal{G}_\alpha &= L_\alpha^c \cap \{m_0 \le 0,\ m_0 + e_\alpha > 0\}, \\ \mathcal{B}_\alpha &= L_\alpha^c \cap \{m_0 > 0,\ m_0 + e_\alpha \le 0\}. \end{aligned} \tag{35}$$

The power change is $P_{\mathrm{M}}(\mathcal{G}_\alpha) - P_{\mathrm{M}}(\mathcal{B}_\alpha)$; the FPR change is $P_{\mathrm{H}}(\mathcal{G}_\alpha) - P_{\mathrm{H}}(\mathcal{B}_\alpha)$.

Proof. The inclusive rank rule is equivalent to $S_j > \tau_j$, including ties. Subtracting thresholds gives the margins; only crossings outside L_α change the complete decision. Taking expectations gives the power and FPR differences. For $k_+ = 0$, neither upper rule rejects. □

C Implementation, Calibration and Evaluation Details

This appendix supplements Algorithm 1 and the evaluation protocol in Section 4.1.

Fixed settings. Parameters follow Section 3, with score floor $\varepsilon = 10^{-9}$ and input limit $L = 2048$. Sensitivity studies identify changes from these defaults.

C.1 Numerical Implementation

Precision and reference support. Tokenization retains at most L tokens, including special tokens, and requires at least two, giving $1 \le n \le L - 1$. Forward passes use FP16; log probabilities and moments use FP32. We compute $\sigma_{jt}^2 = \mathbb{E}_{q_{jt}} \ell_t^2 - \mu_{jt}^2$, clipping negative round-off values to zero. Cached moments yield FP64 scores with variance floor 10^{-18}; the lower score retains its $1/n$ factors when applying this floor.

The nucleus includes the token crossing c. Full support is retained at $c = 1$ or if rounding prevents a crossing. Reported GPU sorts did not enforce stable ordering of equal probabilities. Observed tokens retain their original $\ell_t(x_t)$. Under Equation (6), weights sum to one when capped entropy sums to at least ε_w, and to less otherwise. Zero entropy receives zero weight, including when the cap floor is active; all-zero entropies give $S_+ = 0$. Calibration and queries use identical conventions.

Computational cost. One OPT forward pass supplies both branches. Sorting costs $O(n|\mathcal{V}| \log |\mathcal{V}|)$, moments cost $O(n|\mathcal{V}|)$, and weight normalization costs $O(n)$. Sorting calibration arrays costs $O(m \log m)$ once, enabling $O(\log m)$ tail counts per query instead of Algorithm 1's $O(m)$ direct counts. Binary decisions can use Equation (36).

Two-model Falcon variant. Falcon-7B supplies q_{0t}; Falcon-7B-Instruct supplies ℓ_t over their aligned vocabulary. The single-model identity $q_{0t} = p_t$ therefore does not apply. Shaping uses Falcon-7B logits; entropy uses q_{0t} and $\log q_{0t}$ in Equation (5). Moments and observed residuals use Falcon-7B-Instruct log probabilities. Other formulas are unchanged, with two model evaluations and a separate set of calibration scores.

C.2 Tail Thresholds and Marginal Error Control

For the inclusive tail p-values in Equation (9), define $k_+ = \lfloor (m+1)\alpha\gamma \rfloor$ and $k_- = \lfloor (m+1)\alpha(1-\gamma) \rfloor$. For positive indices, Equation (10) is equivalent to

$$S_+(\mathbf{x}) > c^+_{[k_+]} \quad \text{or} \quad S_-(\mathbf{x}) < c^-_{(k_-)}, \tag{36}$$

where $c^+_{[k]}$ and $c^-_{(k)}$ are the kth largest upper and kth smallest lower calibration scores. Strict inequalities account for ties; a zero-index branch never rejects. Default indices are 4 at nominal 1% and 20 at nominal 5%.

Proposition 5 (Marginal false-positive control)**.** *Suppose the score functions and γ are fixed independently of the human calibration and query sample, whose $m + 1$ documents are exchangeable. Then Equation* (10) *rejects a human document with probability at most α, allowing dependence between branches.*

Proof. Within each branch, exchangeability makes the query's randomly tie-broken rank uniform on $\{1, \dots, m+1\}$. Counting all ties inclusively can only increase its rank, so $P(p_\pm \le u) \le \lfloor (m+1)u \rfloor / (m+1) \le u$. The union bound gives $P(\varphi_\alpha = 1) \le \alpha\gamma + \alpha(1-\gamma) = \alpha$. □

The guarantee is marginal over bank and query sampling. Experimental sample reuse and observed FPRs are described in Appendices C.3 and F.

C.3 Evaluation Protocol

Parameter selection and calibration. We selected T, c, κ by AUROC on the calibration partition in Section 4.1, then fixed them for evaluation. Its 800 humans, 200 per dataset, also supplied empir-

ical tail p-values. This reuse differs from Proposition 5's assumption of independently fixed score functions. Table 4 reports observed evaluation FPRs.

Data sources and generation. Table 9 lists human sources from MAGE (Li et al., 2024). Generators were GLM-4-9B-0414 (Team GLM et al., 2024),[1] Falcon-7B (Almazrouei et al., 2023), Llama-2-7B (Touvron et al., 2023), and OPT-2.7B/6.7B (Zhang et al., 2022). Detection scores human and machine continuations after removing their shared generation prefix. Generation, retry and filtering policies follow Appendix A, including MarkLLM implementations (Pan et al., 2024).

Table 9: Human sources retained for evaluation after overlap exclusion. Each source is paired with five generators and eleven generation conditions. Dataset labels match Table 10.

Label	Dataset and source	Humans
CMV	ChangeMyView (Tan et al., 2016)	200
CNN/DM	CNN/DailyMail (Hermann et al., 2015; Nallapati et al., 2016)	200
DialogSum	DialogSum (Chen et al., 2021b)	162
ELI5	ELI5 (Fan et al., 2019)	200
HellaSwag	HellaSwag (Zellers et al., 2019)	200
IMDb	IMDb (Maas et al., 2011)	199
PubMedQA	PubMedQA (Jin et al., 2019)	200
ROCStories	ROCStories (Mostafazadeh et al., 2016)	200
SciXGen	SciXGen (Chen et al., 2021a)	200
TLDR News	TLDR News (MAGE collection; Li et al., 2024)	200
WritingPrompts	WritingPrompts (Fan et al., 2018)	199
Yelp	Yelp (Zhang et al., 2015)	200
Total unique human sources		2,360

TLDR News source: `https://huggingface.co/datasets/JulesBelveze/tldr_news`.

Watermark implementation settings. DIP used $\alpha_{\mathrm{D}} = 0.45$. SynthID used internal/outer temperatures 0.7/1.0; EXP's accepted retries used 0.7, 0.8, 0.9 or 1.0. ITSEdit used 200 steps of a 256-scalar stream, zero offset and a 0.9999 clamp. State handling and generation-law effects are described in Appendix A.

Source separation and counting. One calibration Writing source overlapped a test WritingPrompts source at 89.47% five-gram containment. Excluding that evaluation source and its generations left 2,360 evaluation and 800 calibration humans with disjoint identities and text hashes. Evaluation contains 11,800 generator-paired human entries and 129,800 machine texts. Pooled FPR counts each human once, retaining the first lexicographic generator/row record.

Baseline scores and precision. Model pairs and input limits follow Section 4.1. Likelihood and Entropy average observed log probabilities and predictive entropy, respectively. Rank and Log-Rank negate mean token rank and log rank. Directions are fixed across conditions; entropy normalization does not change ranking.

AdaDetectGPT and Binoculars use Falcon-7B/Falcon-7B-Instruct, a 2,048-token limit, BF16 forward passes and FP32 calculations, marked † in tables. Ada retains released B-spline coefficients, using Falcon-7B for sampling and Instruct for scoring (Zhou et al., 2025). Binoculars uses the same pair as observer/performer, extends its default 512-token limit, and negates its ratio to orient larger scores toward machine text (Hans et al., 2024). Fast-DetectGPT and our detector use FP16 forward passes and FP32 moments; NTS uses BF16 passes and an algebraically equivalent FP32 moment calculation.

Metrics and uncertainty. Metrics follow Section 4.1. Power, FPR and FP subscripts 1/5 denote nominal total levels, not observed FPR. FPR is the observed human rejection rate; FP is its count. TPR@1%FPR and TPR@5%FPR instead constrain test-human FPR. Calibration-induced ties can make raw-score and empirical-tail AUROC differ.

[1]Release: `https://huggingface.co/zai-org/GLM-4-9B-0414`.

Paired bootstrap intervals resample the twelve evaluation datasets, retaining every source, generator and condition within each sampled dataset. We use 20,000 resamples for baseline comparisons and 10,000 for two-sided-control and reference-shaping comparisons. Intervals describe dataset variation conditional on fixed scoring rules and the calibration bank.

D Extended Detection Results and Two-Sided Controls

D.1 Results by Dataset

Table 10 extends Table 2 to individual datasets under the configurations in Table 1. Our rows rank by J; baselines retain their original score directions. Sources appear in Appendix C.3.

Table 10: Mean AUROC by dataset, averaging five generators and eleven generation conditions. Panels (a) and (b) cover six datasets each. Bold: best in each column, including displayed ties; red: AUROC below 0.5. †: as in Table 1. CMV and CNN/DM denote ChangeMyView and CNN/DailyMail.

(a)

Detector	CMV	CNN/DM	DialogSum	ELI5	HellaSwag	IMDb
Likelihood	0.4815	0.2767	0.3300	0.4831	0.7448	0.5633
Entropy	0.7184	**0.8801**	0.7547	0.6331	0.1477	0.5696
Rank	0.5880	0.4124	0.4159	0.6074	0.6819	0.6244
Log-Rank	0.4963	0.2964	0.3412	0.4983	0.7568	0.5889
Fast-DetectGPT	0.6906	0.5417	0.5758	0.6491	0.6025	0.6939
AdaDetectGPT†	0.6860	0.5485	0.5646	0.6238	0.5920	0.6946
Binoculars†	0.7040	0.5824	0.5614	0.6577	0.4444	0.6919
NTS	0.8604	0.7760	0.7109	0.8128	0.4611	0.8267
Ours (OPT)	0.8152	0.7762	**0.8340**	0.8078	**0.8387**	0.8241
Ours (Falcon pair)	**0.8647**	0.8100	0.8325	**0.8557**	0.7939	**0.8565**

(b)

Detector	PubMedQA	ROCStories	SciXGen	TLDR News	Writing Prompts	Yelp
Likelihood	0.5684	0.7780	0.5130	0.6294	0.4540	0.5313
Entropy	0.4736	0.1257	0.6130	0.3425	0.7121	0.5683
Rank	0.5473	0.6826	0.5585	0.5963	0.5673	0.5597
Log-Rank	0.5853	0.7829	0.5343	0.6415	0.4780	0.5424
Fast-DetectGPT	0.6406	0.6718	0.6529	0.6138	0.6579	0.6746
AdaDetectGPT†	0.6394	0.6565	0.6454	0.6064	0.6639	0.6762
Binoculars†	0.6328	0.5254	0.6547	0.5466	0.6950	0.6661
NTS	0.7490	0.5444	0.7859	0.6585	**0.8700**	0.7531
Ours (OPT)	0.8147	**0.8644**	0.7971	**0.8592**	0.8212	0.8104
Ours (Falcon pair)	**0.8175**	0.8398	**0.8042**	0.8226	0.8578	**0.8247**

D.2 Two-Sided Controls Using One Score

We substitute a baseline score S_B, $B \in \{\mathrm{FD}, \mathrm{Ada}\}$, into both tails of Equation (9). At $\gamma = 1/2$, the resulting p-values give

$$J_B = -\log[2\min\{p_{B,+}, p_{B,-}\}], \qquad \varphi_{B,\alpha} = \mathbf{1}\{p_{B,+} \le \alpha/2 \text{ or } p_{B,-} \le \alpha/2\}. \tag{37}$$

Calibration accommodates unequal human tails, so J_B generally ranks differently from $|S_B|$; negative values are retained.

FD uses our cached lower score, the analytical discrepancy (1) with its numerical floor (Bao et al., 2024). Comparing FD with our test changes only the upper reference and weights, holding the lower score, sources, allocation and decision rule fixed. Both OPT and Falcon are evaluated. Ada uses

the Falcon pair and released B-spline coefficients in both tails (Zhou et al., 2025); its comparison changes both scoring constructions. Configurations and source counting follow Appendix C.3.

Table 4 reports overall performance; Table 11 gives paired dataset-bootstrap intervals. Power contrasts share a nominal budget, with observed FPRs reported separately. Table 12 supplies mechanism-level results for these matched controls, complementing the original-direction baselines in Table 3.

Table 11: Our asymmetric test minus the named two-sided control. Power differences are percentage points at nominal total 1%. Bold: positive gains with 95% confidence intervals above zero. Intervals condition on the fixed rules and human bank. †: as in Table 4.

Proxy / control	ΔAUROC [95% CI]	ΔPower_1 [95% CI]
OPT / FD	**+0.0410** [0.0317, 0.0521]	**+3.27** [1.52, 5.16]
Falcon pair / FD	**+0.0549** [0.0438, 0.0680]	**+4.28** [2.95, 5.79]
Falcon pair / Ada†	**+0.0744** [0.0644, 0.0862]	**+4.88** [3.22, 6.67]

Table 12: AUROC of two-sided scores by condition, averaging 60 evaluation settings per row. FD and Ada use one original score for both tails; our columns use separate upper and lower statistics. Bold: best within each proxy. All columns use the same calibration sources and $\gamma = 1/2$. †: as in Table 4.

	OPT		Falcon pair		
Condition	FD	Ours	FD	Ada†	Ours
Unwatermarked	0.8537	**0.8803**	0.8828	0.8655	**0.9053**
KGW	0.6907	**0.7880**	0.7249	0.6884	**0.8204**
SWEET	0.7049	**0.7782**	0.7472	0.7090	**0.8215**
Unigram	0.6612	**0.7450**	0.6471	0.6194	**0.7532**
EXP	0.9860	**0.9882**	0.9902	0.9912	**0.9942**
SynthID	**0.9805**	0.9802	0.9840	0.9852	**0.9857**
DIP	0.7805	**0.8180**	0.8126	0.7849	**0.8488**
ITSEdit	0.5706	**0.6217**	0.5391	0.5210	**0.6274**
CGW	0.8811	**0.9077**	0.8356	0.8186	**0.8907**
HeavyWater	0.8976	**0.9114**	0.8385	0.8230	**0.8790**
SemStamp	0.5828	**0.6222**	0.5429	0.5234	**0.6221**

E Directional Weighting and Reference-Shaping Ablations

These analyses supplement Section 5. For cap and reference sensitivity, all settings use the full calibration and evaluation cohorts, with $m = 800$, $\gamma = 1/2$ and the lower branch fixed. Each setting recomputes upper calibration scores; parameters other than those listed remain at their defaults. Sensitivity tables bold the best displayed metric, including ties: FPR is minimized; AUROC, CVaR and power are maximized. Bold parameters identify the highest mean AUROC; red marks AUROC or CVaR below 0.5.

E.1 Upper-Tail Mean Residual Differences

Write $d_{jt} = \ell_t(x_t) - \mu_{jt}$, $u_t = 1/n$, and let Cov_U denote covariance under uniform sampling of a document's positions. For unwatermarked machine text, define the unstandardized difference

$$\mathcal{D}^+(w) = \mathbb{E}_{\mathrm{M}} \sum_t w_t d_{1t} - \mathbb{E}_{\mathrm{H}} \sum_t w_t d_{1t}. \tag{38}$$

If $\sum_t w_t = 1$, then

$$\mathcal{D}^+(w) - \mathcal{D}^+(u) = \mathbb{E}_{\mathrm{M}}\, \mathrm{Cov}_U(nw_t, d_{1t}) - \mathbb{E}_{\mathrm{H}}\, \mathrm{Cov}_U(nw_t, d_{1t}). \tag{39}$$

Entropy weighting enlarged this difference in every unwatermarked evaluation setting: OPT's mean increased from 0.3854 to 0.5251 nats; Falcon's from 0.3043 to 0.4325. These measurements precede variance normalization and complement Equation (29).

E.2 Lower-Tail Mean Residual Differences

For each watermark condition in Table 6, reverse the subtraction to match the lower-tail orientation:

$$\mathcal{D}^{-}(w) = \mathbb{E}_{\mathrm{H}} \sum_t w_t d_{0t} - \mathbb{E}_{\mathrm{M}} \sum_t w_t d_{0t}. \tag{40}$$

Normalized weights satisfy $\sum_t w_t d_{0t} - n^{-1} \sum_t d_{0t} = \mathrm{Cov}_U(nw_t, d_{0t})$, so the separation lost by replacing uniform with entropy weights is

$$\Delta_{U-E} := \mathcal{D}^{-}(u) - \mathcal{D}^{-}(w) = \mathbb{E}_{\mathrm{M}} \mathrm{Cov}_U(nw_t, d_{0t}) - \mathbb{E}_{\mathrm{H}} \mathrm{Cov}_U(nw_t, d_{0t}). \tag{41}$$

Positive values in Table 6 indicate reduced lower-tail separation, accounting for changes to both human and machine residuals.

Reference-center effect. Equation (33) isolates center changes at the original scale. Let ρ_{repair} and ρ_{damage} be mean magnitudes of increases and decreases in pairwise ranking credit, with half-credit for ties. Table 6 reports the net loss $\mathcal{L}_{\mathrm{center}} = \rho_{\mathrm{damage}} - \rho_{\mathrm{repair}}$. These losses support retaining the original lower reference center.

E.3 Entropy-Cap Sensitivity

We varied κ with matched OPT forward passes. Effective position count n_{eff} is defined in Appendix F.2.

Table 13: Full-cohort OPT sensitivity to the entropy-cap multiplier κ, with $T = 1.2$, $c = 0.8$, $m = 800$ and $\gamma = 1/2$. The last column gives median n_{eff}/n over unique evaluation humans. All rows, including the default 1.5, use the same newly computed upper moments. Power/FPR are percentages.

κ	AUROC	CVaR_{10}	Power_1	FPR_1	Power_5	FPR_5	n_{eff}/n
1	0.8111	0.5419	38.07	**1.14**	53.69	6.78	0.872
1.25	0.8176	0.5451	38.99	**1.14**	54.46	6.65	0.842
1.5	0.8219	0.5473	40.32	1.31	**54.90**	6.65	0.816
1.75	**0.8232**	0.5474	**41.17**	1.48	54.88	**6.57**	0.798
2	0.8228	0.5472	40.93	1.40	54.69	6.82	0.789
3	0.8218	**0.5475**	39.85	1.40	54.10	6.65	0.783

Tightening the cap from 1.5 to 1 increased the median human effective-position fraction from 0.816 to 0.872, but reduced AUROC from 0.8219 to 0.8111. Relative to 1.5, paired 95% intervals for AUROC changes were $[-0.0161, -0.0062]$ at 1 and $[-0.0065, -0.0022]$ at 1.25. Intervals at 1.75, 2 and 3 included zero, using 10,000 dataset resamples with the bank fixed.

The cap and local reference studies use a matched recomputed default, differing from the main cache by -0.000007 AUROC. Human decisions agree at nominal 1%. At 5%, recomputed power/FPR are 54.90/6.65% versus 54.86/6.69%, with 157 versus 158 human false positives. Contrasts within these studies use the recomputed default.

E.4 Reference-Shaping Gains and Sensitivity

E.4.1 Calibrated Gains and Boundary Crossings

Table 14 expands Section 5.2's comparison with entropy weights fixed. Benefits vary by mechanism: shaping improved joint AUROC for CGW/HeavyWater and ITSEdit/SemStamp, while reducing unwatermarked AUROC for both proxies.

At nominal 1%, shaping added eight and removed six human rejections for OPT, and nineteen and nine for Falcon. Thus observed FPR increased for both. The paired intervals in Section 5.2 use 10,000 dataset resamples with configurations and bank fixed.

Table 14: Shaped minus original upper reference with entropy weights fixed. ΔA_+ and ΔA are AUROC changes for $J_+ = -\log p_+$ and J, averaged across evaluation settings. Six upward WMs: KGW, SWEET, Unigram, EXP, SynthID and DIP. Bold: positive changes.

Proxy	Conditions	ΔA_+	ΔA
OPT	Unwatermarked	-0.0039	-0.0109
OPT	Six upward WMs	-0.0029	-0.0061
OPT	CGW / HeavyWater	**+0.0503**	**+0.0061**
OPT	ITSEdit / SemStamp	**+0.0661**	**+0.0329**
Falcon pair	Unwatermarked	-0.0004	-0.0028
Falcon pair	Six upward WMs	**+0.0022**	**+0.0021**
Falcon pair	CGW / HeavyWater	**+0.0784**	**+0.0234**
Falcon pair	ITSEdit / SemStamp	**+0.1021**	**+0.0722**

E.4.2 Reference Temperature and Nucleus Mass

The factorial comparison varies $T \in \{1, 1.2\}$ and $c \in \{0.8, 1\}$ for both proxies; $c = 1$ retains full support. Entropy weights remain based on the original reference.

Table 15: Complete-test sensitivity to reference temperature T and nucleus mass c, with $\kappa = 1.5$, $\gamma = 1/2$ and $m = 800$. Power and observed unique-human FPR are percentages at nominal total 1% and 5%. All eight settings use the complete evaluation cohort.

Proxy	T	c	AUROC	CVaR_{10}	Power_1	FPR_1	Power_5	FPR_5
OPT	1.0	0.8	0.6924	0.1358	**41.19**	15.47	**55.35**	26.27
OPT	1.0	1.0	0.8191	0.5184	40.87	**1.23**	54.96	**6.53**
OPT	**1.2**	**0.8**	**0.8219**	**0.5473**	40.32	1.31	54.86	6.69
OPT	1.2	1.0	0.6662	0.1439	37.12	22.33	53.67	30.38
Falcon pair	1.0	0.8	0.7103	0.1385	39.47	15.72	54.87	25.89
Falcon pair	1.0	1.0	0.8134	0.5182	40.89	**1.86**	**56.37**	7.54
Falcon pair	**1.2**	**0.8**	**0.8317**	**0.5835**	**44.54**	2.29	55.83	**6.53**
Falcon pair	1.2	1.0	0.6823	0.1481	42.82	24.24	56.09	29.83

Changing temperature or truncation alone reduced OPT AUROC to 0.6662 or 0.6924 and increased nominal-1% FPR to 22.33% or 15.47%, respectively. Falcon showed the same ordering. Raising T and reducing c move the single-model reference center in opposing directions (Proposition 3); variance and recalibrated thresholds also affect decisions.

Local sensitivity. Table 16 changes T or c by 0.1, using matched new forward passes and the recomputed default described in Appendix E.3.

Table 16: Local OPT sensitivity around $(T, c) = (1.2, 0.8)$, with $\kappa = 1.5$, $\gamma = 1/2$ and $m = 800$. Each nondefault row changes only T or c by 0.1. All settings use matched forward passes over the full cohort; power and observed FPR are percentages.

T	c	AUROC	CVaR_{10}	Power_1	FPR_1	Power_5	FPR_5
1.1	0.8	0.7690	0.3326	**42.26**	5.17	**54.90**	13.05
1.2	0.7	0.7837	0.4028	40.67	2.84	54.30	10.42
1.2	**0.8**	**0.8219**	**0.5473**	40.32	**1.31**	**54.90**	**6.65**
1.2	0.9	0.7591	0.3471	39.00	8.47	54.59	16.82
1.3	0.8	0.7429	0.3175	36.06	9.41	53.66	18.52

All four perturbations reduced AUROC and increased nominal-1% FPR. At $(T, c) = (1.3, 0.8)$, AUROC fell by 0.0791, with paired 95% interval $[-0.1540, -0.0107]$ from 10,000 dataset resamples. These results show that performance depends on the joint reference settings.

E.5 Attributing Test Decisions to the Two Tails

Let $U_\alpha = \{p_+ \le \alpha\gamma\}$ and use L_α from Proposition 4. At the complete test's branch thresholds,

$$P(U_\alpha \cup L_\alpha) = P(U_\alpha \setminus L_\alpha) + P(L_\alpha \setminus U_\alpha) + P(U_\alpha \cap L_\alpha). \tag{42}$$

Adding the lower event preserves upper detections at the allocated budget $\alpha\gamma$. Ranking can still change because $J = \max\{-\log p_+, -\log p_-\} - \log 2$ at $\gamma = 1/2$.

Table 17 uses the main scores and 800-human bank. Each condition averages 60 settings; All machine averages 660. Human rates count 2,360 unique sources once. Union rates reproduce Table 4; exclusive events quantify each branch's additional detections. At nominal 1%, upper/lower-only false positives numbered 15/16 for OPT and 48/6 for Falcon.

Table 17: Attribution of test rejections at $\alpha = 1\%$ and $\gamma = 1/2$ (percent). Upper only/lower only are exclusive events; both is the intersection of the two rejection events. Bold: dominant exclusive branch per proxy (both contributions for All machine), and the better union rate across proxies (power maximized, human FPR minimized). Entries are rounded independently.

Nominal total budget 1%								
	OPT				Falcon pair			
Population	Upper only	Lower only	Both	Union	Upper only	Lower only	Both	Union
Unwatermarked	**48.22**	0.05	0.00	48.27	**59.02**	0.00	0.00	**59.02**
KGW	**25.83**	0.07	0.00	25.90	**36.36**	< 0.01	0.00	**36.37**
SWEET	**25.08**	0.05	0.00	25.13	**36.52**	0.03	0.00	**36.55**
Unigram	**22.22**	0.26	0.00	22.49	**25.98**	0.02	0.00	**26.00**
EXP	**91.24**	< 0.01	0.00	91.25	**96.43**	0.00	0.00	**96.43**
SynthID	**86.48**	0.00	0.00	86.48	**91.26**	0.00	0.00	**91.26**
DIP	**31.93**	0.03	0.00	31.95	**43.45**	< 0.01	0.00	**43.46**
ITSEdit	1.64	**4.23**	0.00	5.87	**4.61**	1.44	0.00	**6.05**
CGW	0.65	**49.62**	0.00	**50.27**	2.96	**41.87**	0.00	44.83
HeavyWater	0.08	**50.60**	0.00	**50.67**	0.74	**44.33**	0.00	45.07
SemStamp	1.00	**4.28**	0.00	**5.27**	**3.04**	1.89	0.00	4.93
All machine	**30.40**	**9.93**	0.00	40.32	**36.40**	**8.14**	0.00	**44.54**
Human (FPR)	0.64	0.68	0.00	**1.31**	2.03	0.25	0.00	2.29

F Calibration Transfer and Length Sensitivity

These analyses extend the reliability results in Section 5.3.

Domain variation. Beyond the pooled FPRs in Table 4, worst-domain rates for OPT/Falcon were 2.50%/4.02% at nominal 1% and 11.50%/10.00% at nominal 5%.

F.1 Error Allocation and Calibration-Bank Size

Allocation sensitivity. With the bank and cohort fixed, $\gamma \in \{0.25, 0.5, 0.75\}$ gave AUROCs 0.7919/0.8219/0.8297 for OPT and 0.8161/0.8317/0.8270 for Falcon. Increasing the upper share improved unwatermarked AUROC for both. We retain the preset $\gamma = 1/2$.

Bank size. The minimum tail p-value is $1/(m+1)$. At equal allocation and $\alpha = 0.01$, rejection therefore requires $m \ge 199$. Bank sizes 199, 400 and 800 give order indices 1, 2 and 4.

We drew 20 stratified subsets without replacement for each $m \in \{199, 400\}$, using common seeds across proxies and domain counts differing by at most one. The full bank was evaluated once. Table 18 reports variation across subsets; domain representation affects transfer alongside p-value resolution.

Table 18: Sensitivity to calibration-bank size m at $\gamma = 1/2$ and $\alpha = 1\%$. Entries are median [minimum, maximum] percentages over 20 subsets for $m < 800$; the full-bank row is a single result.

Proxy	m	Power$_1$	FPR$_1$
OPT	199	39.13 [27.48, 45.15]	**0.97** [0.17, 4.32]
OPT	400	39.32 [36.80, 45.56]	1.14 [0.72, 3.05]
OPT	800	**40.32**	1.31
Falcon pair	199	43.39 [27.04, 49.74]	**2.22** [0.08, 4.58]
Falcon pair	400	44.40 [40.87, 47.07]	2.29 [1.82, 3.22]
Falcon pair	800	**44.54**	2.29

F.2 Effective Length and Weight Concentration

For nonnegative normalized weights, define

$$n_{\text{eff}} = \frac{1}{\sum_t w_t^2}, \qquad N_j = \sum_t w_t[\ell_t(x_t) - \mu_{jt}], \qquad \bar{\sigma}_j^2 = \frac{\sum_t w_t^2 \sigma_{jt}^2}{\sum_t w_t^2}. \tag{43}$$

Cauchy–Schwarz gives $1 \le n_{\text{eff}} \le n$, with the upper bound attained exactly by uniform weights. With the score floor inactive,

$$S_j = \frac{N_j \sqrt{n_{\text{eff}}}}{\sqrt{\bar{\sigma}_j^2}}. \tag{44}$$

For entropy weights, this also requires an inactive normalization floor. With only one nonzero-entropy position, capping leaves $n_{\text{eff}} = 1$ and $n_{\text{eff}}/n = 1/n$.

With persistent human mean b for N_j and stable $\bar{\sigma}_j^2$, mean scores can grow approximately as $b\sqrt{n_{\text{eff}}}/\sqrt{\bar{\sigma}_j^2}$. Variance normalization therefore allows accumulation of both machine evidence and human–reference mismatch.

F.3 Natural-Length Stratification of the Complete Test

Table 19 reports detection performance and human FPR by prediction count n. The bins were fixed before computing stratum results. Scores, parameters and the pooled bank remain unchanged. Tokenizers may assign the same text to different bins across proxies.

Stratum AUROC averages settings containing both classes in that bin; supported settings can therefore differ between bins. Power pools all machine texts in the bin, including settings without a same-bin human, rather than weighting settings equally as in the main results. Human FPR counts each source once and uses descriptive Wilson 95% intervals.

Observed effects. OPT power at nominal 1% rose from 9.40% at 1–64 positions to 45.14% at 129–256 positions. Human false positives were more frequent above 512 positions than in the shortest bin. These patterns are consistent with Equation (44), but bins differ in datasets and generation conditions, and the longest bin contains few machine texts. They combine length and composition effects.

Table 19: Length-stratified detection and calibration. (a) AUROC averages settings containing both classes; power pools machine texts. (b) FP/humans counts false positives over unique humans; FPRs include Wilson 95% intervals, and n_{eff}/n gives the human median. Power/FPR are percentages; subscripts 1/5 denote nominal total levels.

(a) Detection performance

Positions n	Humans	Machines	Settings	AUROC	Power$_1$	Power$_5$
			OPT			
1–64	581	5949	313	0.6711	9.40	20.86
65–128	493	17654	407	0.7590	23.28	40.65
129–256	597	105290	495	**0.8414**	**45.14**	**59.33**
257–512	431	885	100	0.7546	35.93	54.46
513–2047	258	22	12	0.6612	31.82	50.00
			Falcon pair			
1–64	580	5836	312	0.6438	12.94	22.96
65–128	486	16990	407	0.7704	24.94	38.69
129–256	591	106252	495	**0.8573**	**49.65**	**60.57**
257–512	428	708	97	0.7919	21.75	34.32
513–2047	275	14	9	0.5378	7.14	14.29

(b) Human false positives

Positions n	FP$_1$/ humans	FPR$_1$ 95% CI	FP$_5$/ humans	FPR$_5$ 95% CI	n_{eff}/n
		OPT			
1–64	0/581	**0.00** [0.00, 0.66]	17/581	**2.93** [1.83, 4.64]	0.854
65–128	6/493	1.22 [0.56, 2.63]	29/493	5.88 [4.13, 8.32]	0.812
129–256	11/597	1.84 [1.03, 3.27]	43/597	7.20 [5.39, 9.56]	0.808
257–512	7/431	1.62 [0.79, 3.31]	37/431	8.58 [6.29, 11.61]	0.794
513–2047	7/258	2.71 [1.32, 5.49]	32/258	12.40 [8.92, 16.99]	0.804
		Falcon pair			
1–64	11/580	1.90 [1.06, 3.36]	35/580	**6.03** [4.37, 8.28]	0.842
65–128	8/486	**1.65** [0.84, 3.21]	33/486	6.79 [4.88, 9.38]	0.798
129–256	13/591	2.20 [1.29, 3.73]	37/591	6.26 [4.58, 8.51]	0.790
257–512	12/428	2.80 [1.61, 4.84]	27/428	6.31 [4.37, 9.02]	0.779
513–2047	10/275	3.64 [1.99, 6.56]	22/275	8.00 [5.34, 11.82]	0.784