CATS: POWER-OPTIMIZED CONFORMAL SELECTION FOR AI-TEXT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Detecting machine-generated text is critical for manuscript review, academic-integrity enforcement, and content moderation, where each accusation carries direct consequences for the flagged author. However, existing methods predominantly rely on heuristic detection scores without statistical guarantees; even when calibrated to control the false positive rate (FPR), such fixed-threshold rules leave the composition of the selected set unconstrained, risking an output dominated by falsely accused human texts. In this work, we formalize machine-generated text detection as a candidate selection problem and propose Conformal AI-Text Selection (**CATS**), a distribution-free framework that controls the false discovery rate (FDR) in finite samples, ensuring the expected proportion of human texts within the selected set stays below a user-specified level q. Specifically, CATS estimates the selection threshold from pooled calibration and candidate data through a permutation-equivariant map, and learns a score adjustment by maximizing a local surrogate for selection power. Theoretically, we prove that this pooled permutation equivariance preserves the super-uniformity of null conformal p-values, thereby guaranteeing finite-sample FDR control. Extensive experiments across ten benchmark datasets and nine detection scores demonstrate that CATS consistently maintains valid FDR control below the target level q while improving selection power over the unadjusted baseline, with an average relative gain of 38% in the main comparison. Our code is available at https://anonymous.4open.science/r/cats-46D4.

1 INTRODUCTION

The rapid development of large language models (LLMs) has made high-quality text generation widely accessible across education, publishing, and online platforms (Brown et al., 2020). However, this accessibility also facilitates misuse, including undisclosed machine-written coursework and synthetic content produced at scale (Zellers et al., 2019b; Weidinger et al., 2021). In response, organizations now routinely screen incoming text for machine generation, using the outcomes to reject manuscripts, initiate academic-integrity proceedings, or restrict content (Liang et al., 2023; Cotton et al., 2024). Given that false accusations carry severe disciplinary consequences for innocent authors and expose institutions to contentious disputes and reputational damage, detection methods must provide credible evidence for texts flagged as machine-generated.

Prior work has developed a range of works to detect machine-generated text, including statistics-based detectors (Gehrmann et al., 2019; Mitchell et al., 2023; Bao et al., 2024) and supervised detectors (Solaiman et al., 2019; Hu et al., 2023). However, these methods merely output detection scores without theoretical guarantees. While some works (Zhu et al., 2025a; Zhou et al., 2026b) calibrate a threshold on a validation set to bound the false positive rate (FPR), this control applies on average over all human texts rather than over the texts the rule actually selects. In realistic pools where human-authored text constitutes the vast majority, a low FPR can still produce a flagged set consisting mostly of misidentified human texts, leaving the flagged set entirely unreliable.

In this paper, we formulate the identification of machine-generated text as a candidate selection problem that controls the false discovery rate (FDR), defined as the expected proportion of human texts within the selected set. Since detection scores are designed for global discriminability rather than for selection power under a fixed error budget, we propose CATS (Conformal AI-Text Selec-

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.

tion), which locates the operating selection boundary through a permutation-equivariant map that pools calibration and candidate data symmetrically. It then learns a score adjustment to the detector score by maximizing a local surrogate for selection power. Calibrating these adjusted scores against the human calibration sample yields conformal p-values, which are then passed to the Benjamini–Hochberg procedure at the user-specified level q. Theoretically, we show that this pooled permutation equivariance preserves the super-uniformity of the null conformal p-values, so that CATS controls the FDR at the user-specified level q with finite samples.

Extensive experiments across ten benchmark datasets and nine detection scores demonstrate the effectiveness of CATS for risk-controlled machine-generated text detection. Across all evaluated settings, CATS reliably controls the FDR at the user-specified target level q, providing strong empirical support for our theoretical guarantees. Furthermore, CATS consistently improves selection power over the baseline across datasets, target levels, and detection scores, achieving an average relative gain of 38% in the main comparisons. For example, at target level $q = 0.10$ when detecting text generated by GPT-3.5 Turbo (Brown et al., 2020; Ouyang et al., 2022) on ChangeMyView (Tan et al., 2016), our method increases the selection power of Fast-DetectGPT from 0.138 to 0.450, and that of Rank from 0.286 to 0.732. Overall, these results show that our method maintains valid FDR control in high-stakes auditing while substantially enhancing selection power across all settings.

Our contributions are as follows.

1. We formulate the identification of machine-generated text as a candidate selection problem that controls the false discovery rate (FDR), directly bounding the expected proportion of human texts within the selected set below the target level q.
2. We propose Conformal AI-Text Selection (**CATS**), which locates the selection threshold through a pooled permutation-equivariant map and learns a score adjustment by maximizing a local surrogate for selection power. We prove that applying the Storey–BH procedure to the resulting conformal p-values controls the FDR at the target level q in finite samples.
3. We conduct extensive experiments across ten benchmark datasets and nine detection scores, demonstrating that CATS reliably maintains valid FDR control below the target level q while consistently improving selection power over baseline detection scores.

2 Background

Machine-generated text detection. Given an input text x, machine-generated text detection aims to determine whether x was authored by a human or generated by a language model. Let $L(x) \in \{0, 1\}$ denote the true provenance of x, where $L(x) = 1$ indicates machine-generated text and $L(x) = 0$ indicates human text. Machine-generated text detection is conventionally formulated as a binary classification problem, in which a detector maps each text to a real-valued detector score $T(x) \in \mathbb{R}$. To help understand, we provide a concrete example of a detection score, Fast-DetectGPT (Bao et al., 2024) below:

$$T_{\mathrm{fd}}(x) = \frac{\log p_\theta(x) - \widetilde{\mu}(x)}{\widetilde{\sigma}(x)}, \tag{1}$$

where p_θ is a scoring language model, and $\widetilde{\mu}(x)$ and $\widetilde{\sigma}(x)$ denote the mean and standard deviation of the log-likelihood under p_θ over alternative texts drawn from a sampling model conditioned on the token prefixes of x. A large positive value means the tokens of x are more probable under p_θ than those of the sampled alternatives.

Supervised detectors instead train a classifier on labeled data, as in the OpenAI GPT-2 detector (Solaiman et al., 2019) and RADAR (Hu et al., 2023). In Section 4.1 we evaluate nine scores from both families. Given such a score, an auditor decides by level-set estimation:

$$\widehat{L}(x) = \mathbb{1}\{T(x) \geq c\}. \tag{2}$$

In practice, some supervised detectors retain the classifier's default boundary, flagging any text whose estimated machine probability exceeds 0.5 (Solaiman et al., 2019). Alternatively, recent works (Zhu et al., 2025a; Zhou et al., 2026b) leverage a reference set to control the FPR.

However, auditing a candidate pool is inherently a multiple testing problem. Because downstream actions such as rejection, content restriction, or disciplinary sanctions are applied to the selected candidates, an output set dominated by human texts produces mostly wrongful accusations. Rather than dividing false selections by all human texts in the pool as the FPR does, the auditor needs to control their expected proportion within the selected set, which is the false discovery rate (FDR).

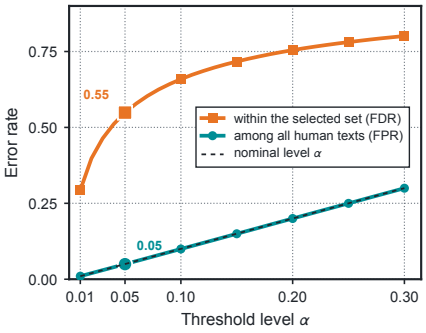


Figure 1: Falsely selected human texts, counted among all human texts (FPR) and within the selected set (FDR). Curves average 1,000 draws of the pool.

To see how large this gap can grow, consider a university auditing 2,000 student essays of which 10% are machine-generated. Figure 1 compares the FPR against the proportion of human texts among all selected texts, where the detector threshold is chosen so that the FPR is exactly α. In particular, at the level $\alpha = 0.05$, the FPR is exactly 0.05 as designed; but 90 of the 1,800 human essays still cross the threshold. Of the 164 essays flagged in total, 90 are human-written—54.9% of the selected set. Therefore, controlling the FPR does not yield a selected set that can be treated as machine-generated. This motivates us to bound the expected proportion of human texts within the selected set, which is the focus of the next section.

3 METHOD

3.1 PROBLEM FORMULATION

To control the proportion of human texts within the selected set used for downstream actions such as disciplinary sanctions or manuscript rejection, we formalize machine-generated text detection as a candidate selection problem. Suppose an auditor has a reference set $\mathcal{D}_{\text{ref}}$ of texts with known provenance and a candidate pool $\mathcal{D}_{\text{test}} = \{X_j^{\text{test}}\}_{j=1}^m$ whose true provenance is unknown, we formulate this task as a multiple hypothesis testing problem:

$$H_{0,j} : X_j^{\text{test}} \sim P_0 \qquad \text{against} \qquad H_{1,j} : X_j^{\text{test}} \not\sim P_0, \tag{3}$$

where P_0 denote the distribution of human text and P_1 that of machine-generated text. Rejecting $H_{0,j}$ indicates that X_j^{test} is flagged as machine-generated text and its index j is included in the selected set $\mathcal{S} \subseteq \{1, \dots, m\}$. To control wrongful accusations, we evaluate the selected set using the false discovery rate (FDR), defined as the expectation of the false discovery proportion (FDP):

$$\text{FDR}(\mathcal{S}) = \mathbb{E}[\text{FDP}(\mathcal{S})], \qquad \text{FDP}(\mathcal{S}) = \frac{\sum_{j=1}^m \mathbb{1}\{j \in \mathcal{S},\, H_{0,j} \text{ is true}\}}{1 \vee |\mathcal{S}|} \tag{4}$$

where $a \vee b := \max\{a, b\}$. The primary requirement is to ensure that $\text{FDR}(\mathcal{S}) \leq q$ for a user-specified level $q \in (0, 1)$. Subject to this rigorous error control, the auditor also seeks to select as many machine-generated texts as possible. We quantify this utility by the **power**:

$$\text{Power}(\mathcal{S}) = \mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{1}\{j \in \mathcal{S},\, H_{1,j} \text{ is true}\}}{1 \vee \sum_{j=1}^m \mathbb{1}\{H_{1,j} \text{ is true}\}}\right]. \tag{5}$$

An effective method should control the FDR below q while maximizing the selection power.

3.2 CATS: CONFORMAL AI-TEXT SELECTION

To turn a detection score into evidence that can support the FDR guarantee, we first calibrate it against a calibration set $\mathcal{D}_{\text{cal}} = \{X_i^{\text{cal}}\}_{i=1}^{n_{\text{cal}}}$ carved from the human portion of $\mathcal{D}_{\text{ref}}$, yielding the conformal p-value for each $H_{0,j}$:

$$p_j(s) = \frac{1 + \sum_{i=1}^{n_{\text{cal}}} \mathbb{1}\{s(X_i^{\text{cal}}) \geq s(X_j^{\text{test}})\}}{n_{\text{cal}} + 1}. \tag{6}$$

where $s(\cdot)$ is obtained by orienting and standardizing the raw detector score T so that larger values indicate stronger evidence of machine generation. Specifically, CATS uses an adjusted score s_q that

targets selection power at the user-specified level q, whose derivation and optimization are detailed in Section 3.3, with full algorithmic details deferred to Appendix C.1.

Under $H_{0,j}$ and the scores are exchangeable, p_j is super-uniform (Vovk et al., 2005; Bates et al., 2023). Ordering $p_{(1)} \le \cdots \le p_{(m)}$, we adopt Storey–BH procedure (Storey et al., 2004):

$$\widehat{\mathcal{S}}_q = \left\{ j : p_j \le \frac{q k_q^*}{m \widehat{\pi}_0^+} \right\}, \quad k_q^* = \max\left\{ k : p_{(k)} \le \frac{qk}{m \widehat{\pi}_0^+} \right\}, \quad \widehat{\pi}_0^+ = \frac{1 + \sum_{j=1}^m \mathbb{1}\{p_j \ge \lambda\}}{m(1-\lambda)}, \tag{7}$$

where $\widehat{\pi}_0^+$ conservatively estimates the proportion of human texts in the candidate pool. Before detailing the construction of the adjusted score s_q, we first motivate this adjustment below.

The AUROC–power gap. However, existing supervised detectors (Solaiman et al., 2019; Hu et al., 2023; Zhou et al., 2026a) are trained for global separation and evaluated by AUROC, which is misaligned with the more critical objective of selection power in the low-q regime. Table 1 illustrates how this objective mismatch degrades selection power: despite achieving an AUROC above 0.91, RADAR yields a vanishingly small selection power of 0.030 at $q = 0.20$ (and 0.005 at $q = 0.10$). Strikingly, an unoptimized zero-shot statistic (predictive entropy) with a much lower AUROC (0.797) recovers six times as many machine texts (0.181) at the same target. To achieve high selection power, we therefore optimize the score directly for the low-q regime, as developed in the following subsection.

Table 1: Baseline selection power on ChangeMyView with GPT-3.5 Turbo.

Detection score	AUROC	Power		
		$q = 0.10$	$q = 0.20$	$q = 0.30$
RADAR	0.917	0.005	0.030	0.137
AdaDetectGPT	0.829	0.002	0.030	0.209
Entropy	0.797	0.019	0.181	0.334

3.3 Adjusted score

Score parameterization and the original power objective. We construct the adjusted score from T by first standardizing $T(x)$ into $s_0(x) = (T(x) - \widehat{\mu}_0)/\widehat{\sigma}_0$, where $\widehat{\mu}_0$ and $\widehat{\sigma}_0$ are the sample mean and standard deviation of T estimated on human reference texts. Let $\phi(x) \in \mathbb{R}^d$ be a frozen auxiliary feature representation of text x. We consider the linear score adjustment

$$s_\beta(x) = s_0(x) + \phi(x)^\top \beta, \tag{8}$$

where $\beta \in \mathbb{R}^d$ is the parameter vector to be optimized. For machine texts $X \sim P_1$, the selection power (Equation (5)) is therefore governed by the threshold-crossing probability:

$$\mathcal{P}_{\tau_q}(\beta) = \mathbb{P}_{X \sim P_1}\big(s_\beta(X) \ge c_q(\beta)\big) = \mathbb{E}_{X \sim P_1}\big[\mathbb{1}\{s_\beta(X) - c_q(\beta) \ge 0\}\big]. \tag{9}$$

Because the discrete indicator provides no gradient signal, directly maximizing this objective is infeasible. Instead, we locate the effective decision cutoff on the pooled data and construct a smooth surrogate around this boundary, as detailed below.

Selection boundary under drift. Under Storey–BH at level q, candidate selection is governed by whether an adjusted score clears an effective, data-dependent cutoff determined by the candidate pool $\mathcal{D}_{\text{test}}$. However, estimating this threshold from $\mathcal{D}_{\text{test}}$ alone introduces an asymmetric dependence that breaks the exchangeability of null scores between $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{test}}$. To maintain pooled equivariance and preserve null-score exchangeability, CATS estimates the selection threshold by combining the final calibration texts and candidate texts into an unlabeled multiset $\mathcal{U} = \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}} = \{Z_i\}_{i=1}^N$. We then construct the corresponding p-values using independent pilot human texts $\mathcal{D}^0_{\text{pilot}}$ from reference data:

$$a_i = \frac{1 + \sum_{\ell=1}^{n_{\text{pilot}}} \mathbb{1}\{s_0(X_\ell^{\text{pilot}}) \ge s_0(Z_i)\}}{n_{\text{pilot}} + 1}, \qquad X_\ell^{\text{pilot}} \in \mathcal{D}^0_{\text{pilot}}. \tag{10}$$

The pooled tail count $R_U(t) = \sum_{i=1}^N \mathbb{1}\{a_i \le t\}$ aggregates candidate observations with the n_{cal} calibration texts, whose pilot p-values have the known marginal null CDF $G_{0,\text{pilot}}(t) = \lfloor (n_{\text{pilot}} +$

$1)t\rfloor/(n_{\text{pilot}}+1)$. Subtracting their expected contribution, which uses no individual role label, gives the debiased candidate tail estimate

$$\widehat{G}_C(t) = \frac{R_U(t) - n_{\text{cal}} G_{0,\text{pilot}}(t)}{m}. \tag{11}$$

Following the Storey–BH procedure, we determine $\widehat{\tau}_q$ as the largest threshold satisfying the target FDR level, and convert it to a baseline score cutoff via the pilot human quantile function:

$$\widehat{\tau}_q = \max\{t \in \mathcal{T}(\mathcal{U}) : \widehat{\pi}_{0,C}\, t \le q\widehat{G}_C(t)\}, \qquad \widehat{c}_q = \widehat{F}^{-1}_{0,\text{pilot}}(1-\widehat{\tau}_q). \tag{12}$$

Here $\mathcal{T}(\mathcal{U})$ is a finite search grid formed by the pooled pilot p-values $\{a_i\}$, and $\widehat{\pi}_{0,C}$ is a Storey estimator for the candidate human proportion.

While $\widehat{c}_q$ anchors the baseline boundary at $\beta = 0$, adjusting the score via β alters the human score distribution and causes the cutoff to drift. To preserve the baseline human tail error budget $\tau_q = \Pr_{X\sim P_0}\{s_0(X) \ge c_q\}$ under score adjustment $s_\beta(X)$, we define the drifting cutoff $c_q(\beta)$ by $\Pr_{X\sim P_0}\{s_\beta(X) \ge c_q(\beta)\} = \tau_q$ with $c_q(0) = c_q$. A first-order Taylor expansion at $\beta = 0$ isolates the drift rate:

$$c_q(\beta) = c_q + \nabla_\beta c_q(0)^\top \beta + O(\|\beta\|^2). \tag{13}$$

The following lemma characterizes the gradient $\nabla_\beta c_q(0)$ explicitly.

Lemma 1. *Suppose $s_0(X)$ admits a continuous density f_0 strictly positive at c_q, and $\mathbb{E}_{P_0}[\phi(X) \mid s_0(X) = y]$ is continuous in y near c_q. Then $c_q(\beta)$ is differentiable at $\beta = 0$ with*

$$\nabla_\beta c_q(0) = \mathbb{E}_{P_0}\big[\phi(X) \mid s_0(X) = c_q\big].$$

The corresponding proof is deferred to Appendix B.1. Expanding the difference between the adjusted score and the cutoff around $\beta = 0$ gives

$$s_\beta(X) - c_q(\beta) = s_0(X) + \phi(X)^\top\beta - \big(c_q + \nabla_\beta c_q(0)^\top\beta\big) + O(\|\beta\|^2) \approx s_0(X) - c_q + \psi(X)^\top\beta, \tag{14}$$

where $\psi(x) = \phi(x) - \nabla_\beta c_q(0)$, which forms the basis for the surrogate below.

Smooth power surrogate. To obtain informative gradient signals, we relax the step indicator using the standard logistic sigmoid $\sigma_{\text{sig}}(t) = 1/(1+e^{-t})$ with bandwidth $h > 0$, yielding the differentiable population surrogate

$$\Pi_q(\beta) = \mathbb{E}_{X\sim P_1}\left[\sigma_{\text{sig}}\left(\frac{s_0(X) - c_q + \psi(X)^\top\beta}{h}\right)\right]. \tag{15}$$

Drawing development human texts $\mathcal{D}^0_{\text{dev}}$ and machine texts $\mathcal{D}^1_{\text{dev}}$ from reference data $\mathcal{D}_{\text{ref}}$, we estimate ψ on $\mathcal{D}^0_{\text{dev}}$ via Gaussian kernel regression at $\widehat{c}_q$ as specified in Appendix C.1, yielding $\widehat{\psi}$. The empirical surrogate on the N_1 machine texts $\{X_i\}_{i=1}^{N_1} \subset \mathcal{D}^1_{\text{dev}}$ then becomes

$$\widehat{\Pi}_q(\beta) = \frac{1}{N_1}\sum_{i=1}^{N_1}\sigma_{\text{sig}}\left(u_i + \frac{\widehat{\psi}(X_i)^\top\beta}{h}\right), \qquad u_i = \frac{s_0(X_i) - \widehat{c}_q}{h}. \tag{16}$$

Because the drift expansion in Equation (13) is local to $\beta = 0$, we approximate $\widehat{\Pi}_q(\beta)$ by the concave quadratic model $\widehat{\Pi}_q(\beta) \approx \widehat{\Pi}_q(0) + \widehat{g}_q^\top\beta - \frac{1}{2}\beta^\top\widehat{A}_q\beta$, with

$$\widehat{g}_q = \frac{1}{N_1}\sum_{i=1}^{N_1}\kappa_i\widehat{\psi}(X_i), \qquad \widehat{A}_q = \frac{1}{N_1}\sum_{i=1}^{N_1}\kappa_i\widehat{\psi}(X_i)\widehat{\psi}(X_i)^\top + \gamma I_d, \tag{17}$$

where $\kappa_i = \frac{1}{h}\sigma_{\text{sig}}(u_i)[1-\sigma_{\text{sig}}(u_i)]$ weights machine texts near the operating boundary $\widehat{c}_q$, and the ridge parameter $\gamma > 0$ guarantees that $\widehat{A}_q$ is strictly positive definite. To keep the adjustment within the valid neighborhood of the local approximation, we enforce an ellipsoidal trust-region constraint $\beta^\top\widehat{A}_q\beta \le r^2$ ($r > 0$). Because the ellipsoidal metric aligns with the quadratic curvature, this constrained maximization admits the exact closed-form clipped Newton step:

$$\widehat{\beta}_q = \arg\max_{\beta:\,\beta^\top\widehat{A}_q\beta\le r^2}\left\{\widehat{g}_q^\top\beta - \frac{1}{2}\beta^\top\widehat{A}_q\beta\right\} = \min\left(1, \frac{r}{\widehat{\nu}_q}\right)\widehat{A}_q^{-1}\widehat{g}_q, \qquad \widehat{\nu}_q = \sqrt{\widehat{g}_q^\top\widehat{A}_q^{-1}\widehat{g}_q}, \tag{18}$$

with $\widehat{\beta}_q = 0$ if $\widehat{\nu}_q = 0$ or if $\widehat{\tau}_q = 0$. The adjusted score is then defined on every text x as

$$s_q(x) = s_0(x) + \phi(x)^\top\widehat{\beta}_q. \tag{19}$$

Substituting s_q into Equation (6) produces the conformal p-values, and applying Equation (7) yields the final selected set $\widehat{\mathcal{S}}_q$. The complete procedure is provided in Algorithm 1 of Appendix C.1.

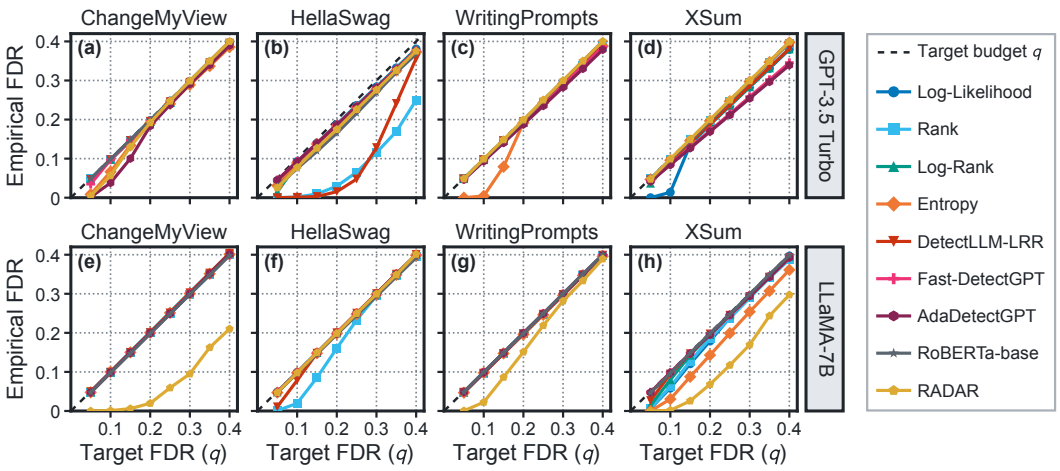


Figure 2: Empirical FDR of our method on four domains with GPT-3.5 Turbo and LLaMA-7B, one curve per detection score. The dashed diagonal marks the target level q.

3.4 Finite-sample FDR control

By construction, CATS adapts to the pooled block $\mathcal{U}$ solely through its multiset while conducting auxiliary estimation on separate reference data, which ensures that the score-generating map satisfies pooled permutation equivariance. Then we establish the finite-sample selection guarantee:

Theorem 1 (Finite-sample FDR control)**.** *Suppose the texts in $\mathcal{D}_{\text{cal}}$ and the human candidates in $\mathcal{D}_{\text{test}}$ are exchangeable, and the adjusted scores $s_q(Z_1), \ldots, s_q(Z_N)$ are almost surely distinct. The selected set $\widehat{\mathcal{S}}_q$ returned by Algorithm 1 satisfies*

$$\text{FDR}(\widehat{\mathcal{S}}_q) \leq q.$$

The proof is provided in Appendix B.2. Crucially, this guarantee is score-agnostic, requiring no distributional or parametric assumptions on the underlying detector scores. Furthermore, statistical validity is decoupled from the power objective, preserving finite-sample FDR control independently of the score adjustment. We empirically confirm this robust control in the next section.

4 Experimental Results

4.1 Setup

Datasets. We evaluate on ten datasets: WritingPrompts (Fan et al., 2018), ChangeMyView (Tan et al., 2016), XSum (Narayan et al., 2018), HellaSwag (Zellers et al., 2019a), Yelp (Zhang et al., 2015), ELI5 (Fan et al., 2019), SQuAD (Rajpurkar et al., 2016), ROCStories (Mostafazadeh et al., 2016), SciXGen (Chen et al., 2021), and TLDR (Belveze, 2022). For each dataset, the human texts and the machine texts produced by GPT-3.5 Turbo and Davinci-003 (Brown et al., 2020; Ouyang et al., 2022) and by LLaMA-7B (Touvron et al., 2023) are taken from the MAGE release (Li et al., 2024). More experimental details are deferred to Appendix C.2.

Detection scores. We consider nine detection scores: Log-Likelihood, Rank, Log-Rank, and Entropy (Gehrmann et al., 2019; Solaiman et al., 2019; Ippolito et al., 2020); RoBERTa-base (Solaiman et al., 2019); DetectLLM-LRR (Su et al., 2023); RADAR (Hu et al., 2023); Fast-DetectGPT (Bao et al., 2024); and AdaDetectGPT (Zhou et al., 2026a).

4.2 Main results

Our method retains FDR control after score adjustment. We first evaluate whether learning the score adjustment preserves strict FDR control. Figure 2 shows the empirical FDR across all

Table 2: Power comparison varying target level and detection scores with GPT-3.5 Turbo. Reported values are the mean with standard error. **Bold** marks the larger value in each pair.

Domain	Detection score	$q = 0.05$		$q = 0.10$		$q = 0.20$		$q = 0.30$	
		Base	Ours	Base	Ours	Base	Ours	Base	Ours
ChangeMyView	Log-Likelihood	0.607±0.002	**0.798±0.001**	0.736±0.001	**0.873±0.000**	0.830±0.001	**0.921±0.000**	0.873±0.000	**0.947±0.000**
	Rank	0.105±0.002	**0.586±0.006**	0.286±0.002	**0.732±0.001**	0.528±0.001	**0.796±0.001**	0.635±0.001	**0.844±0.001**
	Log-Rank	0.551±0.002	**0.775±0.001**	0.700±0.001	**0.849±0.001**	0.808±0.000	**0.911±0.000**	0.857±0.000	**0.941±0.000**
	Entropy	0.002±0.000	**0.011±0.002**	0.019±0.001	**0.149±0.006**	0.181±0.002	**0.559±0.004**	0.334±0.002	**0.682±0.001**
	DetectLLM-LRR	0.205±0.003	**0.314±0.005**	0.442±0.002	**0.664±0.002**	0.620±0.001	**0.796±0.001**	0.722±0.001	**0.866±0.001**
	Fast-DetectGPT	0.024±0.001	**0.092±0.005**	0.138±0.002	**0.450±0.006**	0.393±0.002	**0.745±0.001**	0.618±0.001	**0.851±0.001**
	AdaDetectGPT	0.000±0.000	**0.001±0.000**	0.002±0.000	**0.036±0.003**	0.030±0.001	**0.350±0.009**	0.209±0.004	**0.668±0.005**
	RoBERTa-base	0.223±0.001	**0.439±0.001**	0.369±0.002	**0.481±0.001**	0.554±0.001	**0.603±0.002**	0.708±0.001	**0.880±0.002**
	RADAR	0.000±0.000	**0.012±0.003**	0.005±0.000	**0.162±0.010**	0.030±0.001	**0.663±0.013**	0.137±0.005	**0.943±0.005**
WritingPrompts	Log-Likelihood	0.659±0.002	**0.795±0.001**	0.765±0.001	**0.849±0.000**	0.817±0.000	**0.887±0.000**	0.847±0.000	**0.908±0.000**
	Rank	0.355±0.002	**0.717±0.002**	0.537±0.001	**0.779±0.000**	0.675±0.001	**0.829±0.000**	0.738±0.001	**0.856±0.000**
	Log-Rank	0.649±0.002	**0.795±0.001**	0.755±0.001	**0.847±0.000**	0.812±0.000	**0.884±0.000**	0.846±0.000	**0.904±0.000**
	Entropy	0.000±0.000	0.000±0.000	0.001±0.000	**0.020±0.003**	0.264±0.004	**0.593±0.005**	0.531±0.001	**0.762±0.001**
	DetectLLM-LRR	0.507±0.002	**0.694±0.002**	0.655±0.001	**0.796±0.001**	0.752±0.001	**0.853±0.000**	0.801±0.000	**0.879±0.000**
	Fast-DetectGPT	0.427±0.001	**0.631±0.001**	0.535±0.001	**0.703±0.001**	0.645±0.001	**0.761±0.000**	0.703±0.001	**0.800±0.000**
	AdaDetectGPT	0.171±0.001	**0.431±0.004**	0.269±0.001	**0.590±0.001**	0.468±0.001	**0.683±0.001**	0.586±0.001	**0.742±0.001**
	RoBERTa-base	0.344±0.001	**0.451±0.001**	0.457±0.001	**0.496±0.001**	0.589±0.001	**0.638±0.002**	0.686±0.001	**0.932±0.001**
	RADAR	0.187±0.000	**0.274±0.002**	0.221±0.001	**0.364±0.001**	0.361±0.002	**0.506±0.002**	0.616±0.002	**0.883±0.002**
XSum	Log-Likelihood	0.000±0.000	**0.001±0.000**	0.007±0.002	**0.074±0.006**	0.576±0.001	**0.744±0.001**	0.671±0.001	**0.790±0.000**
	Rank	0.509±0.001	**0.813±0.001**	0.581±0.001	**0.846±0.001**	0.688±0.001	**0.884±0.000**	0.751±0.001	**0.910±0.000**
	Log-Rank	0.087±0.004	**0.280±0.007**	0.420±0.002	**0.642±0.001**	0.618±0.001	**0.759±0.001**	0.686±0.001	**0.797±0.000**
	Entropy	0.333±0.000	**0.336±0.000**	0.334±0.000	**0.371±0.001**	0.362±0.001	**0.713±0.002**	0.478±0.001	**0.847±0.001**
	DetectLLM-LRR	0.422±0.001	**0.576±0.001**	0.517±0.001	**0.669±0.001**	0.624±0.001	**0.746±0.001**	0.678±0.001	**0.783±0.000**
	Fast-DetectGPT	0.461±0.001	**0.587±0.001**	0.541±0.001	**0.634±0.000**	0.599±0.001	**0.655±0.000**	0.624±0.000	**0.664±0.000**
	AdaDetectGPT	0.224±0.003	**0.431±0.004**	0.358±0.001	**0.560±0.001**	0.486±0.001	**0.620±0.001**	0.557±0.001	**0.645±0.000**
	RoBERTa-base	0.467±0.001	**0.543±0.001**	0.564±0.001	**0.590±0.001**	0.686±0.001	**0.706±0.001**	0.770±0.001	**0.884±0.001**
	RADAR	0.114±0.001	**0.221±0.003**	0.146±0.001	**0.444±0.002**	0.198±0.001	**0.585±0.001**	0.561±0.009	**0.899±0.003**

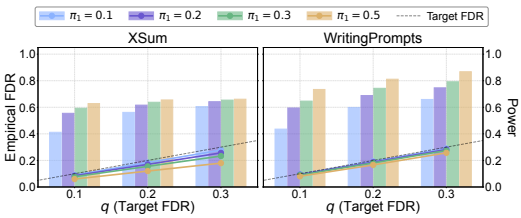


Figure 3: Empirical FDR (lines, left axis) and power (bars, right axis) of our method under varying machine-text fraction π_1 of the candidate pool. The diagonal dashed line is the target FDR.

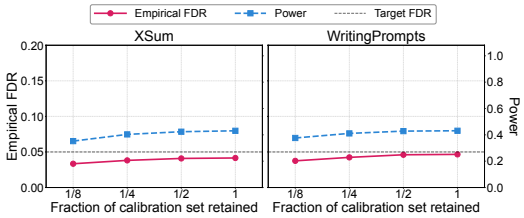


Figure 4: Empirical FDR and power of our method at $q = 0.05$ under varying fraction of the final calibration set retained. The dashed line is the target FDR.

evaluated detection scores with GPT-3.5 Turbo and LLaMA-7B. Across the eight panels arranged by generator rows and datasets columns, each curve presents the empirical FDR of an individual detection score as the target level q varies. Throughout all panels, the empirical FDR across all detection scores is strictly controlled below the dashed diagonal line representing the nominal target level q. These empirical results align with the theoretical guarantee of Theorem 1, confirming that our method reliably maintains valid FDR control across diverse settings.

Our method attains higher power than the baseline. Table 2 compares the selection power of our method against the baseline applying the selection procedure directly to unadjusted detector scores. Across all settings, our method consistently achieves higher selection power than the baseline, yielding an average relative gain of 38% in selection power across the main comparisons. For example, on WritingPrompts with AdaDetectGPT at $q = 0.10$, our method increases the power from 0.269 to 0.590, more than doubling the fraction of machine texts recovered. More experimental results are provided in Appendix D.1 and Appendix D.2.

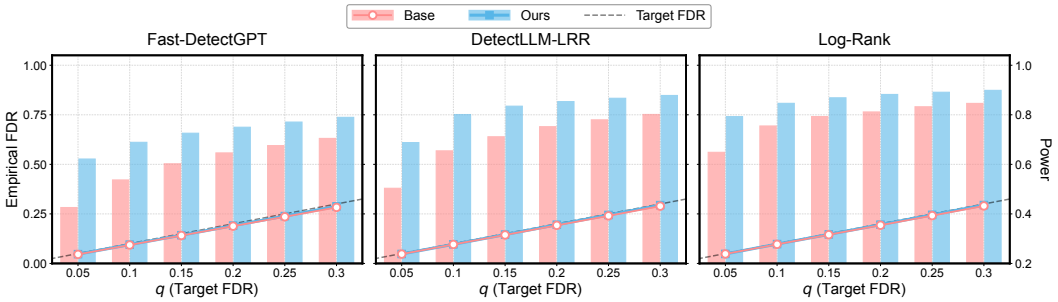


Figure 5: Empirical FDR (lines, left axis) and power (bars, right axis) of our method and the baseline on WritingPrompts, with the score adjustment learned from Davinci-003 machine texts and then applied to a candidate pool generated by GPT-3.5 Turbo. The dashed line is the target FDR.

4.3 Sensitivity analysis

Robustness of error control to variations in the machine-text fraction π_1. In practical audits, the prevalence of machine text π_1 in the candidate pool $\mathcal{D}_{\text{test}}$ is unknown *a priori* and outside the auditor's control. To evaluate the robustness of our method against varying mixture proportions, we sweep $\pi_1 \in \{0.10, 0.20, 0.30, 0.50\}$ across target levels $q \in \{0.10, 0.20, 0.30\}$ on XSum and WritingPrompts with GPT-3.5 Turbo scored by AdaDetectGPT. As shown in Figure 3, the empirical FDR (solid lines) remains strictly bounded by the nominal target diagonal (dashed line) across all evaluated prevalence settings and target levels. This confirms the validity of Theorem 1, which guarantees finite-sample error control regardless of the candidate class balance.

Impact of calibration set size. A critical consideration for deploying conformal selection is sample efficiency: how much labeled human calibration data is required to obtain power gains. To evaluate audit robustness under reference data scarcity, we investigate this by varying the retained fraction of the final human calibration set $\mathcal{D}_{\text{cal}}$ over $\{0.125, 0.25, 0.50, 1.00\}$ on XSum and WritingPrompts, while holding $\mathcal{D}_{\text{test}}$ fixed. Figure 4 presents the resulting empirical FDR and selection power at $q = 0.05$. Throughout all calibration regimes, the empirical FDR stays strictly within the nominal budget, confirming finite-sample guarantee of our method.

Cross-generator transfer of the score adjustment. An auditor may encounter candidate texts produced by a generator not represented in the reference corpus. To assess transferability under such generator mismatch, we learn the score adjustment from Davinci-003 machine texts in the reference data, then deploy it on a candidate pool generated by GPT-3.5 Turbo. Figure 5 presents the empirical FDR and selection power on WritingPrompts for three detection scores across target levels $q \in [0.05, 0.30]$. Across all settings, the empirical FDR tracks the nominal budget line faithfully, and our method consistently outperforms the baseline in selection power. In particular, for Fast-DetectGPT at $q = 0.05$, our method improves selection power from 0.426 to 0.621. These results demonstrate that our method transfers effectively to an unseen generator, preserving reliable error control with higher selection power than the baseline in selection power under generator mismatch.

5 Related Work

Machine-generated text detection. Machine-generated text detection now informs decisions in education, peer review, and online publishing (Solaiman et al., 2019; Weidinger et al., 2021; Liang et al., 2023; Cotton et al., 2024; Liang et al., 2024; Tang et al., 2024; Wu et al., 2025b). Zero-shot detectors score a text with a probability-based statistics (Gehrmann et al., 2019; Solaiman et al., 2019; Ippolito et al., 2020; Su et al., 2023; Hans et al., 2024; Xu et al., 2025; Luo et al., 2026), perturbation-based statistics (Mitchell et al., 2023; Bao et al., 2024; 2025; Yang et al., 2024; Mao et al., 2024; Zhu et al., 2025b; Ma & Wang, 2024), or geometric and ensemble statistics (Tulchinskii et al., 2023; Mireshghallah et al., 2024; Dubois et al., 2025) from a surrogate language model. Supervised detectors fine-tune a classifier on raw text or on extracted features (Solaiman et al., 2019; Zellers et al., 2019b; Verma et al., 2024; Wang et al., 2023; Shi et al., 2024a; Guo et al., 2024a; Kim

et al., 2024; Yu et al., 2024; Lee et al., 2024), train the surrogate model or the scoring statistic itself (Zeng et al., 2024; Chen et al., 2025; Fu et al., 2025; Zhou et al., 2026a;c), or learn kernel two-sample tests (Zhang et al., 2024; Song et al., 2025). Training strategies further include multi-task learning (Li et al., 2026), positive-unlabeled learning (Tian et al., 2024), contrastive learning (Liu et al., 2024; Guo et al., 2024b), adversarial training (Hu et al., 2023; Koike et al., 2024; Li et al., 2025), out-of-distribution detection (Zeng et al., 2025), and easy-to-hard supervision (Wu et al., 2025a). Watermarking offers an alternative when the generator cooperates (Kirchenbauer et al., 2023; 2024; Zhao et al., 2024; Hu et al., 2024; Kuditipudi et al., 2024; Christ et al., 2024; Dathathri et al., 2024), although watermarks can be stolen and spoofed (Jovanovic et al., 2024). Paraphrasing and adversarial rewriting degrade detectors of both kinds (Krishna et al., 2023; Shi et al., 2024b; Wang et al., 2024a; 2025), whether detection remains possible is debated (Sadasivan et al., 2025; Chakraborty et al., 2024), and benchmarks spanning generators, domains, and attacks track detection accuracy (Wang et al., 2024b; He et al., 2024; Li et al., 2024; Dugan et al., 2024; Wu et al., 2024). While some works (Zhu et al., 2025a; Zhou et al., 2026b) calibrate a threshold on a validation set to bound the FPR, and sequential tests bound the Type-I error when texts from one source arrive as a stream (Chen & Wang, 2025), such control averages over human texts and leaves the composition of the selected set unconstrained. In contrast, our method controls the FDR of the selected set for any detector score under black-box score access, and learns a score adjustment that raises selection power while preserving finite-sample FDR control.

Conformal inference and FDR control. Conformal inference provides finite-sample valid evidence under exchangeability (Vovk et al., 2005). In standard conformal prediction, conformal p-values are primarily used to construct prediction sets with guaranteed marginal coverage for individual instances by testing candidate labels (Vovk et al., 2005; Angelopoulos & Bates, 2023). Beyond per-sample prediction, a growing line of work connects conformal inference with selective inference and multiple hypothesis testing: by coupling conformal p-values with the Benjamini–Hochberg (BH) procedure (Benjamini & Hochberg, 1995), one can select candidate subsets under FDR control (Bates et al., 2023; Jin & Candès, 2023; Marandon et al., 2024). Subsequent extensions address more constrained settings, including covariate shift through weighted conformal p-values (Jin & Candès, 2025), multivariate responses through vector-valued conformity scores (Bai et al., 2025), and multiple response conditions through conditional conformity constructions (Hao et al., 2026), as well as the dependence among conformal p-values (Marandon et al., 2024; Gazin et al., 2024). From an efficiency perspective, power can be improved by optimizing conformity scores directly (Bai & Jin, 2024) or by employing clipped score functions (Jin & Candès, 2023); however, both formulations require full exchangeability across calibration and candidate pools drawn from a common joint distribution. In this paper, our method requires exchangeability only between the calibration texts and the human texts in the candidate pool, while maximizing a local surrogate for selection power around the operating selection boundary.

6 Conclusion

In this paper, we formalize machine-generated text detection as a candidate selection problem and propose Conformal AI-Text Selection (**CATS**), a distribution-free framework with finite-sample FDR control. To improve selection power under the target level q, CATS locates the selection boundary and learns a score adjustment by maximizing a surrogate for selection power. Because this adjustment is permutation-equivariant on pooled calibration and candidate data, it preserves null exchangeability and guarantees finite-sample FDR control via conformal p-values. Extensive experiments across diverse benchmarks and detection scores show that our method reliably controls the FDR while consistently improving selection power over the baseline, enabling reliable machine-generated text selection in high-stakes auditing.

Limitations and future work. While CATS achieves consistent power improvements across diverse benchmarks, its score adjustment naturally refines existing detector scores around the operating selection boundary, meaning that the magnitude of power gains reflects the local signal quality provided by the underlying detector. In addition, our FDR guarantee holds marginally over calibration and candidate draws under exchangeability, and extending distribution-free FDR control to candidate pools under distribution shifts presents a natural and valuable direction for future study.

AI USE STATEMENT

We used a coding assistant for routine refactoring and for drafting the plotting and table-generation scripts that produce the figures and tables in this paper, with all outputs reviewed by the authors against the underlying result files. We did not use generative AI to design the method, to derive or check the proofs, to run or interpret the experiments, or to select which results to report. No figure in this paper contains generated imagery. The authors take full responsibility for the content of the manuscript.

REPRODUCIBILITY STATEMENT

The selection procedure is stated in full as Algorithm 1, with hyperparameter specifications and implementation conventions detailed in Appendix C.1. All evaluation corpora are drawn from the public MAGE benchmark (Li et al., 2024), and no text was generated for this work. Data partitioning protocols, role sample sizes (Table 4), and computational costs across the 1000 Monte Carlo repetitions are documented in Appendix C.2. The complete source code, environment specifications, and scripts to reproduce all figures and tables are anonymously available at https://anonymous.4open.science/r/cats-46D4.

REFERENCES

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

Tian Bai and Ying Jin. Optimized conformal selection: Powerful selective inference after conformity score optimization, 2024. URL https://arxiv.org/abs/2411.17983.

Tian Bai, Yue Zhao, Xiang Yu, and Archer Y. Yang. Multivariate conformal selection. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=g2tr7nA4pS.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bpcgcr8E8Z.

Guangsheng Bao, Yanbin Zhao, Juncai He, and Yue Zhang. Glimpse: Enabling white-box methods to use proprietary models for zero-shot llm-generated text detection. In *International Conference on Learning Representations*, volume 2025, pp. 51844–51865, 2025.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1), Feb 2023. ISSN 0090-5364. doi: 10.1214/22-aos2244. URL http://dx.doi.org/10.1214/22-aos2244.

Jules Belveze. TLDR news. Hugging Face dataset, 2022. URL https://huggingface.co/datasets/JulesBelveze/tldr_news.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, Jan 1995. ISSN 1467-9868. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Souradip Chakraborty, Amrit Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. Position: On the possibilities of AI-generated text detection. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=CJbhtpcyGL.

Can Chen and Jun-Kun Wang. Online detection of LLM-generated texts via sequential hypothesis testing by betting. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9231–9276. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chen25bn.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. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.128. URL http://dx.doi.org/10.18653/v1/2021.findings-emnlp.128.

Jiaqi Chen, Xiaoye Zhu, Tianyang Liu, Ying Chen, Chen Xinhui, Yiwen Yuan, Chak Tou Leong, Zuchao Li, Long Tang, Lei Zhang, Chenyu Yan, Guanghao Mei, Jie Zhang, and Lefei Zhang. Imitate before detect: Aligning machine stylistic preference for machine-revised text detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23559–23567, Apr 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i22.34525. URL http://dx.doi.org/10.1609/aaai.v39i22.34525.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1125–1139. PMLR, 2024.

Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in education and teaching international*, 61(2): 228–239, 2024.

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(8035):818–823, Oct 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-08025-4. URL http://dx.doi.org/10.1038/s41586-024-08025-4.

Matthieu Dubois, François Yvon, and Pablo Piantanida. Mosaic: Multiple observers spotting ai content. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24230–24247. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1244. URL http://dx.doi.org/10.18653/v1/2025.findings-acl.1244.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12463–12492. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.674. URL http://dx.doi.org/10.18653/v1/2024.acl-long.674.

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. Association for Computational Linguistics, 2018. doi: 10.18653/v1/p18-1082. URL http://dx.doi.org/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. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1346. URL http://dx.doi.org/10.18653/v1/P19-1346.

Jiachen Fu, Chun-Le Guo, and Chongyi Li. Detectanyllm: Towards generalizable and robust detection of machine-generated text across domains and models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 11229–11238. ACM, Oct 2025. doi: 10.1145/3746027.3754862. URL http://dx.doi.org/10.1145/3746027.3754862.

Ulysse Gazin, Gilles Blanchard, and Etienne Roquain. Transductive conformal inference with adaptive scores. In *International Conference on Artificial Intelligence and Statistics*, pp. 1504–1512. PMLR, 2024.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In Marta R. Costa-jussà and Enrique Alfonseca (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, Florence, Italy, jul 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3019. URL https://aclanthology.org/P19-3019/.

Hanxi Guo, Siyuan Cheng, Xiaolong Jin, Zhuo Zhang, Kaiyuan Zhang, Guanhong Tao, Guangyu Shen, and Xiangyu Zhang. Biscope: Ai-generated text detection by checking memorization of preceding tokens. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 104065–104090. Curran Associates, Inc., 2024a. doi: 10.52202/079017-3307. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/bc808cf2d2444b0abcceca366b771389-Paper-Conference.pdf.

Xun Guo, Shan Zhang, Yongxin He, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. Detective: Detecting ai-generated text via multi-level contrastive learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 88320–88347. Curran Associates, Inc., 2024b. doi: 10.52202/079017-2802. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a117a3cd54b7affad04618c77c2fb18b-Paper-Conference.pdf.

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 *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=axl3FAkpik.

Qingyang Hao, Wenbo Liao, Bingyi Jing, and Hongxin Wei. Multi-condition conformal selection. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 44069–44097, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/490d9bcb9befeacab21d3746924eaa4c-Paper-Conference.pdf.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. Mgtbench: Benchmarking machine-generated text detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pp. 2251–2265. ACM, Dec 2024. doi: 10.1145/3658644.3670344. URL http://dx.doi.org/10.1145/3658644.3670344.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 15077–15095. Curran Associates, Inc., 2023. doi: 10.52202/075280-0662. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/30e15e5941ae0cdab7ef58cc8d59a4ca-Paper-Conference.pdf.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 45408–45436, 2024.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 1808–1822, 2020.

Ying Jin and Emmanuel J Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

Ying Jin and Emmanuel J Candès. Model-free selective inference under covariate shift via weighted conformal ¡i¿p¡/i¿ -values. *Biometrika*, 113(1), Sept 2025. ISSN 1464-3510. doi: 10.1093/biomet/asaf066. URL http://dx.doi.org/10.1093/biomet/asaf066.

Nikola Jovanovic, Robin Staab, and Martin T. Vechev. Watermark stealing in large language models. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22570–22593. PMLR / OpenReview.net, 2024.

Zae Myung Kim, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. Threads of subtlety: Detecting machine-generated texts through discourse motifs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5449–5474. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.298. URL http://dx.doi.org/10.18653/v1/2024.acl-long.298.

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. PMLR, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 49660–49704, 2024.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21258–21266, Mar 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i19.30120. URL http://dx.doi.org/10.1609/aaai.v38i19.30120.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27469–27500. Curran Associates, Inc., 2023. doi: 10.52202/075280-1195. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/575c450013d0e99e4b0ecf82bd1afaa4-Paper-Conference.pdf.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=FpaCL1MO2C.

Hyunseok Lee, Jihoon Tack, and Jinwoo Shin. Remodetect: Reward models recognize aligned llm's generations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 2886–2913. Curran Associates, Inc., 2024. doi: 10.52202/079017-0095. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/058373e239c542eae6cb796e4dc521da-Paper-Conference.pdf.

Chenjun Li, Cheng Wan, and Johannes C. Paetzold. Meld: Multi-task equilibrated learning detector for ai-generated text, 2026. URL https://arxiv.org/abs/2605.06903.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36–53, Bangkok, Thailand, aug 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.3. URL https://aclanthology.org/2024.acl-long.3/.

Yuanfan Li, Zhaohan Zhang, Chengzhengxu Li, Chao Shen, and Xiaoming Liu. Iron sharpens iron: Defending against attacks in machine-generated text detection with adversarial training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pp. 3091–3113. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.155. URL http://dx.doi.org/10.18653/v1/2025.acl-long.155.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7):100779, July 2023. ISSN 2666-3899. doi: 10.1016/j.patter.2023.100779. URL http://dx.doi.org/10.1016/j.patter.2023.100779.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel McFarland, and James Y. Zou. Monitoring AI-modified content at scale: A case study on the impact of chatGPT on AI conference peer reviews. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=bX3J7ho18S.

Shengchao Liu, Xiaoming Liu, Yichen Wang, Zehua Cheng, Chengzhengxu Li, Zhaohan Zhang, Yu Lan, and Chao Shen. Does DetectGPT fully utilize perturbation? bridging selective perturbation to fine-tuned contrastive learning detector would be better. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1874–1889, Bangkok, Thailand, aug 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.103. URL https://aclanthology.org/2024.acl-long.103/.

Haitong Luo, Weiyao Zhang, Suhang Wang, Wenji Zou, Chungang Lin, Xuying Meng, and Yujun Zhang. Specdetect: Simple, fast, and training-free detection of llm-generated text via spectral analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):32356–32364, Mar 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i38.40510. URL http://dx.doi.org/10.1609/aaai.v40i38.40510.

Shixuan Ma and Quan Wang. Zero-shot detection of llm-generated text using token cohesiveness. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17538–17553. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.971. URL http://dx.doi.org/10.18653/v1/2024.emnlp-main.971.

Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. Raidar: generative ai detection via rewriting. In *International Conference on Learning Representations*, volume 2024, pp. 6013–6030, 2024.

Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1), Feb 2024. ISSN 0090-5364. doi: 10.1214/23-aos2338. URL http://dx.doi.org/10.1214/23-aos2338.

Niloofar Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. Smaller language models are better zero-shot machine-generated text detectors. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 278–293. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-short.25. URL http://dx.doi.org/10.18653/v1/2024.eacl-short.25.

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

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. Association for Computational Linguistics, 2016. doi: 10.18653/v1/n16-1098. URL http://dx.doi.org/10.18653/v1/N16-1098.

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. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1206. URL http://dx.doi.org/10.18653/v1/D18-1206.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

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. Association for Computational Linguistics, 2016. doi: 10.18653/v1/d16-1264. URL http://dx.doi.org/10.18653/v1/D16-1264.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? stress testing AI text detectors under various attacks. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=OOgsAZdFOt.

Yuhui Shi, Qiang Sheng, Juan Cao, Hao Mi, Beizhe Hu, and Danding Wang. Ten words only still help: Improving black-box ai-generated text detection via proxy-guided efficient re-sampling. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024, pp. 494–502. International Joint Conferences on Artificial Intelligence Organization, Aug 2024a. doi: 10.24963/ijcai.2024/55. URL http://dx.doi.org/10.24963/ijcai.2024/55.

Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. Red teaming language model detectors with language models. *Transactions of the Association for Computational Linguistics*, 12:174–189, 2024b. ISSN 2307-387X. doi: 10.1162/tacl_a_00639. URL http://dx.doi.org/10.1162/tacl_a_00639.

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

Yiliao Song, Zhenqiao Yuan, Shuhai Zhang, Zhen Fang, Jun Yu, and Feng Liu. Deep kernel relative test for machine-generated text detection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=z9j7wctoGV.

John D Storey, Jonathan E Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1):187–205, 2004.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12395–12412, Singapore, dec 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.827. URL https://aclanthology.org/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*, WWW '16, pp. 613–624,

Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883081. URL https://doi.org/10.1145/2872427.2883081.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated text. *Communications of the ACM*, 67(4):50–59, Mar 2024. ISSN 1557-7317. doi: 10.1145/3624725. URL http://dx.doi.org/10.1145/3624725.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, Qinghua Zhang, Ruifeng Li, Chao Xu, and Yunhe Wang. Multiscale positive-unlabeled detection of ai-generated texts. In *International Conference on Learning Representations*, volume 2024, pp. 34849–34866, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. Intrinsic dimension estimation for robust detection of ai-generated texts. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 39257–39276. Curran Associates, Inc., 2023. doi: 10.52202/075280-1706. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7baa48bc166aa2013d78cbdc15010530-Paper-Conference.pdf.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. Ghostbuster: Detecting text ghostwritten by large language models. 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. 1702–1717. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.95. URL http://dx.doi.org/10.18653/v1/2024.naacl-long.95.

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

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. Seqxgpt: Sentence-level ai-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1144–1156. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.73. URL http://dx.doi.org/10.18653/v1/2023.emnlp-main.73.

Tianchun Wang, Yuanzhou Chen, Zichuan Liu, Zhanwen Chen, Haifeng Chen, Xiang Zhang, and Wei Cheng. Humanizing the machine: Proxy attacks to mislead LLM detectors. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PIpGN5Ko3v.

Yichen Wang, Shangbin Feng, Abe Hou, Xiao Pu, Chao Shen, Xiaoming Liu, Yulia Tsvetkov, and Tianxing He. Stumbling blocks: Stress testing the robustness of machine-generated text detectors under attacks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2894–2925. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.acl-long.160. URL http://dx.doi.org/10.18653/v1/2024.acl-long.160.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1369–1407. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.eacl-long.83. URL http://dx.doi.org/10.18653/v1/2024.eacl-long.83.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models, 2021. URL https://arxiv.org/abs/2112.04359.

Chenwang Wu, Yiu-ming Cheung, Bo Han, and Defu Lian. Advancing machine-generated text detection from an easy to hard supervision perspective. In *Advances in Neural Information Processing Systems 38*, NeurIPS 2025, pp. 166812–166860. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025a. doi: 10.52202/085713-5023. URL http://dx.doi.org/10.52202/085713-5023.

Junchao Wu, Runzhe Zhan, Derek Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S. Chao. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 100369–100401. Curran Associates, Inc., 2024. doi: 10.52202/079017-3186. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b61bdf7e9f64c04ec75a26e781e2ad51-Paper-Datasets_and_Benchmarks_Track.pdf.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338, 2025b. ISSN 1530-9312. doi: 10.1162/coli_a_00549. URL http://dx.doi.org/10.1162/coli_a_00549.

Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. Training-free llm-generated text detection by mining token probability sequences. In *International Conference on Learning Representations*, volume 2025, pp. 19072–19098, 2025.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Ruth Petzold, William Yang Wang, and Haifeng Chen. DNA-GPT: Divergent n-gram analysis for training-free detection of GPT-generated text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Xlayxj2fWp.

Xiao Yu, Kejiang Chen, Qi Yang, Weiming Zhang, and Nenghai Yu. Text fluoroscopy: Detecting LLM-generated text through intrinsic features. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15838–15846, Miami, Florida, USA, nov 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.885. URL https://aclanthology.org/2024.emnlp-main.885/.

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. Association for Computational Linguistics, 2019a. doi: 10.18653/v1/p19-1472. URL http://dx.doi.org/10.18653/v1/P19-1472.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pp. 9051–9062, 2019b. URL https://proceedings.neurips.cc/paper/2019/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html.

Cong Zeng, Shengkun Tang, Xianjun Yang, Yuanzhou Chen, Yiyou Sun, Zhiqiang Xu, Yao Li, Haifeng Chen, Wei Cheng, and Dongkuan Xu. Dald: Improving logits-based detector without logits from black-box llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 54947–54973. Curran Associates, Inc., 2024. doi: 10.52202/079017-1743. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/62e5f22b5dae99ec700be622df4fbe0d-Paper-Conference.pdf.

Cong Zeng, Shengkun Tang, Yuanzhou Chen, Zhiqiang Shen, Wenchao Yu, Xujiang Zhao, Haifeng Chen, Wei Cheng, and Zhiqiang Xu. Human texts are outliers: Detecting llm-generated texts via out-of-distribution detection. In *Advances in Neural Information Processing Systems 38*, NeurIPS 2025, pp. 181393–181423. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025. doi: 10.52202/085713-5457. URL http://dx.doi.org/10.52202/085713-5457.

Shuhai Zhang, Yiliao Song, Jiahao Yang, Yuanqing Li, Bo Han, and Mingkui Tan. Detecting machine-generated texts by multi-population aware optimization for maximum mean discrepancy. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3fEKavFsnv.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, pp. 649–657, 2015.

Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=SsmT8aO45L.

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. *Advances in Neural Information Processing Systems*, 38:89077–89118, 2026a.

Hongyi Zhou, Jin Zhu, Ying Yang, and Chengchun Shi. Detecting llm-generated text with performance guarantees. *arXiv preprint arXiv:2601.06586*, 2026b.

Hongyi Zhou, Jin Zhu, Kai Ye, Ying Yang, Erhan Xu, and Chengchun Shi. Learn-to-distance: Distance learning for detecting LLM-generated text. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=2ZUPeEM3FH.

Xiaowei Zhu, Yubing Ren, Yanan Cao, Xixun Lin, Fang Fang, and Yangxi Li. Reliably bounding false positives: A zero-shot machine-generated text detection framework via multiscaled conformal prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12298–12319, 2025a.

Xiaowei Zhu, Yubing Ren, Fang Fang, Qingfeng Tan, Shi Wang, and Yanan Cao. Dna-detectllm: Unveiling ai-generated text via a dna-inspired mutation-repair paradigm. In *Advances in Neural Information Processing Systems 38*, NeurIPS 2025, pp. 2022–2045. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025b. doi: 10.52202/085713-0065. URL http://dx.doi.org/10.52202/085713-0065.

A TABLE OF NOTATIONS

Table 3: Summary of notations used in the paper.

Notation	Description
x, $L(x)$	A generic text and its true provenance; $L(x) = 1$ marks machine-generated text.
P_0, P_1	The distributions of human text and machine-generated text.
$T(x)$	The raw detector score; larger values are more machine-like.
$\mathcal{D}_{\text{ref}}$	The labeled reference corpus of texts with known provenance.
$\mathcal{D}^a_{\text{pilot}}, \mathcal{D}^a_{\text{dev}}$	The pilot and development data carved from $\mathcal{D}_{\text{ref}}$; together they form $\mathcal{L}$.
$\mathcal{D}_{\text{cal}} = \{X_i^{\text{cal}}\}_{i=1}^{n_{\text{cal}}}$	The final human calibration set.
$\mathcal{D}_{\text{test}} = \{X_j^{\text{test}}\}_{j=1}^{m}$	The candidate pool under audit.
n_{pilot}, π_1	The pilot human sample size and the machine fraction of the candidate pool.
$H_{0,j}, H_{1,j}$	The null that candidate $X_j^{\text{test}} \sim P_0$ is human text, and its alternative.
$\mathcal{H}_0, \mathcal{H}_1$	The index sets of human and machine-generated candidates.
q	The user-specified target FDR level, $q \in (0, 1)$.
$\mathcal{S}$, $\widehat{\mathcal{S}}_q$	A selected set of candidate indices, and the output of Algorithm 1.
$\text{FDP}(\mathcal{S})$, $\text{FDR}(\mathcal{S})$	The realized fraction of human texts in $\mathcal{S}$, and its expectation.
$\text{Power}(\mathcal{S})$	The selection power, i.e., the expected fraction of machine candidates selected.
p_j	The conformal p-value of candidate j; $p_{(k)}$ is the k-th smallest.
λ, $\widehat{\pi}_0^+$	The Storey threshold $K/(n_{\text{cal}}+1)$ and the conservative null-proportion estimator.
k_q^*	The Storey–BH crossing rank at level q.
$s_0(x)$	The standardized baseline score $(T(x) - \widehat{\mu}_0)/\widehat{\sigma}_0$.
$\phi(x) \in \mathbb{R}^d$	The frozen auxiliary feature representation of text x.
β, $s_\beta(x)$	The adjustment parameter vector and the adjusted score $s_0(x) + \phi(x)^\top \beta$.
$\widehat{\beta}_q$, $s_q(x)$	The fitted coefficient and the adjusted score $s_{\widehat{\beta}_q}(x)$ used for selection.
$\mathcal{U} = \{Z_i\}_{i=1}^N$	The pooled multiset $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$, with $N = n_{\text{cal}} + m$.
$\mathcal{I}_0, \mathcal{I}_1$	The null and non-null coordinates of $\mathcal{U}$.
$\mathcal{L}$	The learning-only data $(\mathcal{D}^0_{\text{pilot}}, \mathcal{D}^1_{\text{pilot}}, \mathcal{D}^0_{\text{dev}}, \mathcal{D}^1_{\text{dev}})$.
ω, ωv	A permutation of indices, and the vector v with its coordinates reordered by ω.
$\mathcal{G}_q(\mathcal{U}; \mathcal{L})$	The adjusted score vector $(s_q(Z_1), \ldots, s_q(Z_N))$ assigned to the pooled block.
$S_{\mathcal{I}_0}, S_{\mathcal{I}_1}$	The final adjusted scores of the null and non-null coordinates of $\mathcal{U}$.
$S_i^{\text{cal}}, S_j^{\text{test}}$	The final adjusted scores of calibration text i and candidate j.
a_i	The pilot p-value of Z_i against $\mathcal{D}^0_{\text{pilot}}$.
$R_U(t)$, $\widehat{G}_C(t)$	The pooled tail count and the debiased candidate tail estimate.
$\widehat{\pi}_{0,C}$	The pilot null-proportion estimator with threshold λ_{pilot}.
τ_q, $\widehat{\tau}_q$	The human tail budget $\Pr_{P_0}\{s_0(X) \geq c_q\}$ and its estimate on the grid $\mathcal{T}(\mathcal{U})$.
c_q, $\widehat{c}_q$	The baseline score cutoff at level q and its pilot estimate $\widehat{F}^{-1}_{0,\text{pilot}}(1-\widehat{\tau}_q)$.
$c_q(\beta)$	The drifting cutoff that keeps the human tail probability of s_β at τ_q.
$\psi(x)$, $\widehat{\psi}(x)$	The boundary-centered feature $\phi(x) - \nabla_\beta c_q(0)$ and its estimate.
σ_{sig}, h	The logistic sigmoid and its smoothing bandwidth.
$\Pi_q(\beta)$, $\widehat{\Pi}_q(\beta)$	The smooth power surrogate and its version over N_1 machine development texts.
u_i, κ_i	The standardized boundary margin of machine text X_i and its local weight.
$\widehat{g}_q$, $\widehat{A}_q$	The local power direction and the curvature matrix with ridge γ.
r, $\widehat{\nu}_q$	The trust-region radius and the curvature-metric length $\sqrt{\widehat{g}_q^\top \widehat{A}_q^{-1} \widehat{g}_q}$.

B PROOFS

B.1 THE PROOF OF LEMMA 1

Proof. For a human text $X \sim P_0$, denote the baseline standardized detector score by $Y = s_0(X) \in \mathbb{R}$ and the auxiliary feature vector by $V = \phi(X) \in \mathbb{R}^d$, so that the linear score adjustment in Equation (8) is $s_\beta(X) = Y + V^\top \beta$. Recall from Section 3.3 that the drifting human cutoff $c_q(\beta)$ preserves the baseline human tail error budget:

$$\Pr_{X \sim P_0}\big\{s_\beta(X) \geq c_q(\beta)\big\} = \tau_q, \qquad c_q(0) = c_q,$$

where $\tau_q = \Pr_{X \sim P_0}\{s_0(X) \geq c_q\} \in (0,1)$. Fix an arbitrary direction $v \in \mathbb{R}^d$ with $\|v\| = 1$, and consider the one-dimensional perturbation $\beta(t) = tv$ for $t \in (-\epsilon, \epsilon)$. For any $a \in \mathbb{R}$, define the perturbed cumulative distribution function under P_0 as

$$F_{tv}(a) := \Pr_0\big\{s_{tv}(X) \leq a\big\} = \Pr_0\big\{Y + tV^\top v \leq a\big\}. \tag{20}$$

The fixed-tail probability constraint is equivalently expressed as $F_{tv}(c_q(tv)) = 1 - \tau_q$. Differentiating both sides with respect to t at $t = 0$ via the multivariable chain rule yields

$$\begin{aligned} 0 &= \frac{d}{dt} F_{tv}(c_q(tv))\Big|_{t=0} \\ &= \frac{\partial}{\partial t} F_{tv}(a)\Big|_{\substack{t=0 \\ a=c_q}} + \frac{\partial}{\partial a} F_0(a)\Big|_{a=c_q} \frac{d}{dt} c_q(tv)\Big|_{t=0} \\ &= \frac{\partial}{\partial t} F_{tv}(c_q)\Big|_{t=0} + f_{0,Y}(c_q)\, \frac{d}{dt} c_q(tv)\Big|_{t=0}, \end{aligned} \tag{21}$$

where $F_0(a) = \Pr_0\{Y \leq a\}$ is the baseline marginal CDF of Y with continuous density $f_{0,Y}(c_q) > 0$. To evaluate the partial derivative $\frac{\partial}{\partial t} F_{tv}(c_q)\big|_{t=0}$, we condition on V using the law of total expectation:

$$F_{tv}(c_q) = \mathbb{E}_0\big[\Pr_0\{Y \leq c_q - tV^\top v \mid V\}\big] = \mathbb{E}_0\big[F_{Y|V}(c_q - tV^\top v \mid V)\big], \tag{22}$$

where $F_{Y|V}(y \mid u) = \Pr_0\{Y \leq y \mid V = u\}$ denotes the conditional CDF of Y given $V = u$. By dominated convergence, exchanging differentiation with integration and applying the chain rule to $F_{Y|V}$ gives

$$\begin{aligned} \frac{\partial}{\partial t} F_{tv}(c_q)\Big|_{t=0} &= \mathbb{E}_0\left[\frac{\partial}{\partial t} F_{Y|V}(c_q - tV^\top v \mid V)\Big|_{t=0}\right] \\ &= \mathbb{E}_0\left[f_{Y|V}(c_q \mid V)(-V^\top v)\right] \\ &= -\mathbb{E}_0\big[f_{Y|V}(c_q \mid V)V\big]^\top v. \end{aligned} \tag{23}$$

Writing this expectation in integral form with the marginal density $f_V(u)$ of V under P_0 and utilizing the joint density factorization $f_{Y|V}(c_q \mid u) f_V(u) = f_{Y,V}(c_q, u)$, we have

$$\begin{aligned} \mathbb{E}_0\big[f_{Y|V}(c_q \mid V)V\big] &= \int_{\mathbb{R}^d} u\, f_{Y|V}(c_q \mid u) f_V(u)\, du \\ &= \int_{\mathbb{R}^d} u\, f_{Y,V}(c_q, u)\, du \\ &= f_{0,Y}(c_q) \left(\frac{\int_{\mathbb{R}^d} u\, f_{Y,V}(c_q, u)\, du}{f_{0,Y}(c_q)}\right) \\ &= f_{0,Y}(c_q) \int_{\mathbb{R}^d} u\, f_{V|Y}(u \mid c_q)\, du \\ &= f_{0,Y}(c_q)\, \mathbb{E}_0[V \mid Y = c_q], \end{aligned} \tag{24}$$

where the fourth line uses Bayes' rule for the conditional density $f_{V|Y}(u \mid c_q) = f_{Y,V}(c_q, u)/f_{0,Y}(c_q)$. Substituting Equation (24) into Equation (23) yields

$$\frac{\partial}{\partial t} F_{tv}(c_q)\Big|_{t=0} = -f_{0,Y}(c_q)\,\mathbb{E}_0[V \mid Y = c_q]^\top v. \tag{25}$$

Substituting Equation (25) back into the chain rule expansion Equation (21), we obtain

$$-f_{0,Y}(c_q)\,\mathbb{E}_0[V \mid Y = c_q]^\top v + f_{0,Y}(c_q)\,\frac{d}{dt} c_q(tv)\Big|_{t=0} = 0. \tag{26}$$

Since $f_{0,Y}(c_q) > 0$, dividing by $f_{0,Y}(c_q)$ isolates the directional derivative:

$$\frac{d}{dt} c_q(tv)\Big|_{t=0} = \mathbb{E}_0[V \mid Y = c_q]^\top v. \tag{27}$$

Because $\frac{d}{dt} c_q(tv)\big|_{t=0} = \nabla_\beta c_q(0)^\top v$ for every direction $v \in \mathbb{R}^d$, we conclude that

$$\nabla_\beta c_q(0) = \mathbb{E}_0[V \mid Y = c_q] = \mathbb{E}_{P_0}\big[\phi(X) \mid s_0(X) = c_q\big], \tag{28}$$

which completes the proof. □

B.2 The proof of Theorem 1

Let $\mathcal{L} = (\mathcal{D}^0_{\text{pilot}}, \mathcal{D}^1_{\text{pilot}}, \mathcal{D}^0_{\text{dev}}, \mathcal{D}^1_{\text{dev}})$ collect the learning-only data carved from $\mathcal{D}_{\text{ref}}$. Let $\mathcal{I}_0$ index the null coordinates of $\mathcal{U}$, namely the n_{cal} calibration texts together with the candidates $j \in \mathcal{H}_0 = \{j : H_{0,j} \text{ is true}\}$, let $\mathcal{I}_1$ index the remaining coordinates, and write $Z_{\mathcal{I}_0}$ and $Z_{\mathcal{I}_1}$ for the corresponding subvectors of $\mathcal{U} = (Z_1, \ldots, Z_N)$. For a permutation ω of indices and a vector v, ωv denotes v with its coordinates reordered by ω. The exchangeability hypothesis of Theorem 1 is used in the following form.

Assumption 1 (Conditional null exchangeability)**.** Conditionally on $(Z_{\mathcal{I}_1}, \mathcal{L})$, the null vector $Z_{\mathcal{I}_0}$ is jointly exchangeable: for every permutation ω of $\mathcal{I}_0$,

$$Z_{\mathcal{I}_0} \mid (Z_{\mathcal{I}_1}, \mathcal{L}) \overset{d}{=} \omega Z_{\mathcal{I}_0} \mid (Z_{\mathcal{I}_1}, \mathcal{L}) \qquad \text{almost surely.}$$

This holds when the human texts of $\mathcal{D}_{\text{ref}}$ and of the candidate pool are i.i.d. draws from P_0 independent of the machine texts, since Algorithm 1 splits $\mathcal{D}_{\text{ref}}$ at random.

Lemma 2 (Storey–BH for adaptive empirical conformal p-values (Marandon et al., 2024, Theorem 3.6 and Corollary 3.7))**.** *Let $S^{\text{cal}}_1, \ldots, S^{\text{cal}}_{n_{\text{cal}}}$ be calibration scores and $S^{\text{test}}_1, \ldots, S^{\text{test}}_m$ candidate scores, with null candidates indexed by $\mathcal{H}_0$. Suppose that*

(i) conditionally on the alternative scores $(S^{\text{test}}_j)_{j \notin \mathcal{H}_0}$, the vector $\big(S^{\text{cal}}_1, \ldots, S^{\text{cal}}_{n_{\text{cal}}}, (S^{\text{test}}_j)_{j \in \mathcal{H}_0}\big)$ is jointly exchangeable, and the $n_{\text{cal}} + m$ scores are almost surely distinct;

(ii) each p_j is the conformal p-value Equation (6) of S^{test}_j against the common calibration scores;

(iii) for a fixed integer $K \in \{2, \ldots, n_{\text{cal}}\}$ and $\lambda = K/(n_{\text{cal}} + 1)$, the null proportion is estimated without truncation at one by

$$\widehat{\pi}^+_0 = \frac{1 + \sum_{j=1}^m \mathbb{1}\{p_j \geq \lambda\}}{m(1 - \lambda)},$$

and $\widehat{\mathcal{S}}_q$ is the Benjamini–Hochberg selected set at level $q/\widehat{\pi}^+_0$, as in Equation (7).

Then $\text{FDR}(\widehat{\mathcal{S}}_q) \leq q$.

Proof of Theorem 1. Fix a realization of $\mathcal{L}$ and a permutation ω of $\{1, \ldots, N\}$, and let $\mathcal{G}_q(\mathcal{U}; \mathcal{L}) = (s_q(Z_1), \ldots, s_q(Z_N))$ be the adjusted scores that Algorithm 1 assigns to $\mathcal{U}$. Given $\mathcal{L}$, the map s_0, the

pilot sample, and all tuning constants are fixed, so Equation (10) is one map applied coordinatewise and $a(\omega\mathcal{U}) = \omega\, a(\mathcal{U})$. Every pooled quantity of Algorithm 1 reads $\mathcal{U}$ only through the multiset $\{a_i\}_{i=1}^N$:

$$R_U(t) = \sum_{i=1}^{N} \mathbb{1}\{a_i \le t\}, \qquad \widehat{G}_C(t) = \frac{R_U(t) - n_{\mathrm{cal}} G_{0,\mathrm{pilot}}(t)}{m},$$
$$\widehat{\tau}_q = \max\{t \in \mathcal{T}(\mathcal{U}) : \widehat{\pi}_{0,C}\, t \le q\, \widehat{G}_C(t)\},$$

and the same holds for $\widehat{\pi}_{0,C}$ in Equation (32) and for the grid $\mathcal{T}(\mathcal{U})$. Hence $\widehat{\tau}_q$ and $\widehat{c}_q = \widehat{F}^{-1}_{0,\mathrm{pilot}}(1 - \widehat{\tau}_q)$ are invariant under ω. The quantities $\widehat{\psi}$, $\widehat{g}_q$, $\widehat{A}_q$, and r are computed on $\mathcal{L}$ and depend on $\mathcal{U}$ only through $\widehat{c}_q$, so $\widehat{\beta}_q$ in Equation (18) is invariant as well, including the branch $\widehat{\beta}_q = 0$ when $\widehat{\tau}_q = 0$. Since $s_q = s_0 + \phi^\top \widehat{\beta}_q$ is then one fixed map applied to every text,

$$\mathcal{G}_q(\omega\mathcal{U}; \mathcal{L}) = \omega\, \mathcal{G}_q(\mathcal{U}; \mathcal{L}). \tag{29}$$

Now let ω permute $\mathcal{I}_0$ and fix $\mathcal{I}_1$, and write $(S_{\mathcal{I}_0}, S_{\mathcal{I}_1}) = \mathcal{G}_q(Z_{\mathcal{I}_0}, Z_{\mathcal{I}_1}; \mathcal{L})$ for the final null and alternative scores. Assumption 1 gives $(Z_{\mathcal{I}_0}, Z_{\mathcal{I}_1}, \mathcal{L}) \overset{d}{=} (\omega Z_{\mathcal{I}_0}, Z_{\mathcal{I}_1}, \mathcal{L})$, and applying $\mathcal{G}_q$ together with Equation (29) yields

$$(S_{\mathcal{I}_0}, S_{\mathcal{I}_1}) \overset{d}{=} \mathcal{G}_q(\omega Z_{\mathcal{I}_0}, Z_{\mathcal{I}_1}; \mathcal{L}) = (\omega S_{\mathcal{I}_0}, S_{\mathcal{I}_1}). \tag{30}$$

The alternative scores $S_{\mathcal{I}_1}$ may depend on the null texts, but only through their multiset, so they are unchanged by ω. For Borel sets A and B, Equation (30) gives

$$\mathbb{E}\big[\mathbb{1}\{S_{\mathcal{I}_1} \in B\} \Pr(S_{\mathcal{I}_0} \in A \mid S_{\mathcal{I}_1})\big] = \mathbb{E}\big[\mathbb{1}\{S_{\mathcal{I}_1} \in B\} \Pr(\omega S_{\mathcal{I}_0} \in A \mid S_{\mathcal{I}_1})\big],$$

so $S_{\mathcal{I}_0}$ is jointly exchangeable conditionally on $S_{\mathcal{I}_1}$.

Fix $j \in \mathcal{H}_0$. The $n_{\mathrm{cal}} + 1$ scores $(S_1^{\mathrm{cal}}, \dots, S_{n_{\mathrm{cal}}}^{\mathrm{cal}}, S_j^{\mathrm{test}})$ are exchangeable by Equation (30) and distinct by hypothesis, so the descending rank of S_j^{test} among them is uniform, and by Equation (6)

$$(n_{\mathrm{cal}} + 1)\, p_j = 1 + \sum_{i=1}^{n_{\mathrm{cal}}} \mathbb{1}\{S_i^{\mathrm{cal}} \ge S_j^{\mathrm{test}}\} \sim \mathrm{Unif}\{1, \dots, n_{\mathrm{cal}} + 1\},$$
$$\Pr(p_j \le t) = \frac{\lfloor (n_{\mathrm{cal}} + 1)t \rfloor}{n_{\mathrm{cal}} + 1} \le t \qquad \text{for every } t \in [0, 1]. \tag{31}$$

Because every p_j is ranked against the same calibration scores, the null p-values are dependent, and the classical finite-sample Storey bound for independent p-values does not apply; Lemma 2 covers this shared-calibration structure. Its condition (i) is Equation (30) together with the distinctness hypothesis of Theorem 1, condition (ii) is Equation (6), and condition (iii) is Equation (7) under the Storey calibration. Hence Lemma 2 gives $\mathrm{FDR}(\widehat{\mathcal{S}}_q) \le q$. □

C Algorithm and Experimental Details

C.1 Algorithm Details

Algorithm 1 summarizes the complete procedure of CATS. The remainder of this subsection records the implementation details and numerical conventions.

Pilot construction. The candidate-side distribution is recovered from the pooled tail count by Equation (11), in which $G_{0,\mathrm{pilot}}$ is the exact distribution of a null pilot p-value under the non-randomized rank formula Equation (10); with randomized pilot values it may be replaced by the identity. The threshold grid is $\mathcal{T}(\mathcal{U}) = (\mathcal{T}_{\mathrm{fixed}} \cup \{0, a_1, \dots, a_N\}) \cap [0, 1 - \epsilon_\tau]$ for a pre-registered deterministic grid $\mathcal{T}_{\mathrm{fixed}}$, where the truncation at $1 - \epsilon_\tau$ avoids endpoint degeneracy; because the data-dependent part enters only through the multiset $\{a_i\}$, pooled symmetry is preserved. Under Equation (12), the feasible set always contains $t = 0$, which serves as the default operating point when no positive threshold satisfies the Storey–BH condition. The null-proportion surrogate applies the same calibration-null correction to the upper tail of the pilot values at a pre-registered λ_{pilot},

$$\widehat{\pi}_{0,C} = \frac{1 + \sum_{i=1}^{N} \mathbb{1}\{a_i > \lambda_{\mathrm{pilot}}\} - n_{\mathrm{cal}} \overline{G}_{0,\mathrm{pilot}}(\lambda_{\mathrm{pilot}})}{m\, \overline{G}_{0,\mathrm{pilot}}(\lambda_{\mathrm{pilot}})}, \quad \overline{G}_{0,\mathrm{pilot}} = 1 - G_{0,\mathrm{pilot}}, \tag{32}$$

Algorithm 1 Conformal AI-Text Selection (CATS)

Require: Detector score T, reference data $\mathcal{D}_{\mathrm{ref}}$, candidate pool $\mathcal{D}_{\mathrm{test}} = \{X_j^{\mathrm{test}}\}_{j=1}^{m}$, target level q, Storey thresholds $\lambda, \lambda_{\mathrm{pilot}}$, bandwidths b, h, ridge γ.

1: Sample $\mathcal{D}^a_{\mathrm{pilot}}, \mathcal{D}^a_{\mathrm{dev}}$ ($a \in \{0,1\}$) and $\mathcal{D}_{\mathrm{cal}} = \{X_i^{\mathrm{cal}}\}_{i=1}^{n_{\mathrm{cal}}}$ from $\mathcal{D}_{\mathrm{ref}}$ without replacement.

2: Estimate the mean $\widehat{\mu}_0$ and standard deviation $\widehat{\sigma}_0$ of T on $\mathcal{D}^0_{\mathrm{pilot}}$.

3: $s_0 \leftarrow \dfrac{T - \widehat{\mu}_0}{\widehat{\sigma}_0}$.

4: Fit the feature map ϕ by PCA on $\mathcal{D}^0_{\mathrm{pilot}} \cup \mathcal{D}^1_{\mathrm{pilot}}$.

5: $\mathcal{U} = \{Z_i\}_{i=1}^{N} \leftarrow \mathcal{D}_{\mathrm{cal}} \cup \mathcal{D}_{\mathrm{test}}$.

6: Discard the role of every observation in $\mathcal{U}$.

7: $a_i \leftarrow \dfrac{1 + \sum_{\ell=1}^{n_{\mathrm{pilot}}} \mathbb{1}\{s_0(X_\ell^{\mathrm{pilot}}) \geq s_0(Z_i)\}}{n_{\mathrm{pilot}} + 1}$ for $i = 1, \ldots, N$.

8: $\widehat{G}_C(t) \leftarrow \dfrac{\sum_{i=1}^{N} \mathbb{1}\{a_i \leq t\} - n_{\mathrm{cal}} G_{0,\mathrm{pilot}}(t)}{m}$ for $t \in \mathcal{T}(\mathcal{U})$.

9: $\widehat{\pi}_{0,C} \leftarrow \dfrac{1 + \sum_{i=1}^{N} \mathbb{1}\{a_i > \lambda_{\mathrm{pilot}}\} - n_{\mathrm{cal}} \overline{G}_{0,\mathrm{pilot}}(\lambda_{\mathrm{pilot}})}{m\, \overline{G}_{0,\mathrm{pilot}}(\lambda_{\mathrm{pilot}})}$.

10: $\widehat{\tau}_q \leftarrow \max\{t \in \mathcal{T}(\mathcal{U}) : \widehat{\pi}_{0,C}\, t \leq q\, \widehat{G}_C(t)\}$.

11: $\widehat{c}_q \leftarrow \widehat{F}^{-1}_{0,\mathrm{pilot}}(1 - \widehat{\tau}_q)$.

12: $w(X) \leftarrow \exp\left(-\dfrac{(s_0(X) - \widehat{c}_q)^2}{2b^2}\right)$ for $X \in \mathcal{D}^0_{\mathrm{dev}}$.

13: $\widehat{\psi}(X_i) \leftarrow \phi(X_i) - \dfrac{\sum_{X \in \mathcal{D}^0_{\mathrm{dev}}} w(X)\phi(X)}{\sum_{X \in \mathcal{D}^0_{\mathrm{dev}}} w(X)}$ for $X_i \in \mathcal{D}^1_{\mathrm{dev}}$.

14: $u_i \leftarrow \dfrac{s_0(X_i) - \widehat{c}_q}{h}$ for $X_i \in \mathcal{D}^1_{\mathrm{dev}}$.

15: $\kappa_i \leftarrow \dfrac{1}{h}\sigma_{\mathrm{sig}}(u_i)[1 - \sigma_{\mathrm{sig}}(u_i)]$ for $X_i \in \mathcal{D}^1_{\mathrm{dev}}$.

16: $\widehat{g}_q \leftarrow \dfrac{1}{N_1}\sum_{i=1}^{N_1} \kappa_i \widehat{\psi}(X_i)$, $\widehat{A}_q \leftarrow \dfrac{1}{N_1}\sum_{i=1}^{N_1} \kappa_i \widehat{\psi}(X_i)\widehat{\psi}(X_i)^\top + \gamma I_d$.

17: Select the trust radius r by development cross-fitting.

18: $\widehat{\nu}_q \leftarrow \sqrt{\widehat{g}_q^\top \widehat{A}_q^{-1} \widehat{g}_q}$.

19: $\widehat{\beta}_q \leftarrow \min(1, r/\widehat{\nu}_q)\, \widehat{A}_q^{-1} \widehat{g}_q$.

20: $s_q(Z_i) \leftarrow s_0(Z_i) + \phi(Z_i)^\top \widehat{\beta}_q$ for $i = 1, \ldots, N$.

21: Restore the roles of $\mathcal{U}$ as $\mathcal{D}_{\mathrm{cal}}$ and $\mathcal{D}_{\mathrm{test}}$.

22: $p_j \leftarrow \dfrac{1 + \sum_{i=1}^{n_{\mathrm{cal}}} \mathbb{1}\{s_q(X_i^{\mathrm{cal}}) \geq s_q(X_j^{\mathrm{test}})\}}{n_{\mathrm{cal}} + 1}$ for $j = 1, \ldots, m$.

23: $\widehat{\pi}_0^+ \leftarrow \dfrac{1 + \sum_{j=1}^{m} \mathbb{1}\{p_j \geq \lambda\}}{m(1 - \lambda)}$.

24: $k_q^* \leftarrow \max\left\{k : p_{(k)} \leq \dfrac{qk}{m\widehat{\pi}_0^+}\right\}$.

25: Return the selected set $\widehat{\mathcal{S}}_q = \left\{j : p_j \leq \dfrac{qk_q^*}{m\widehat{\pi}_0^+}\right\}$.

with $\lambda_{\mathrm{pilot}} = 0.5$ in every reported run. The numerator counts pilot values above λ_{pilot} over the whole pooled block and then removes the expected calibration-null contribution, reading the multiset without consulting any role label.

C.2 Experimental details

Data partitioning and repetitions. For each domain and generator pair, texts are partitioned via stratified random sampling without replacement into the labeled reference corpus $\mathcal{D}_{\mathrm{ref}}$ and the candidate pool $\mathcal{D}_{\mathrm{test}}$, with $\mathcal{D}_{\mathrm{ref}}$ further divided into the pilot, development, and final calibration roles

of Section 3.2. The sample sizes for every role are held constant across repetitions and across all detection scores, as summarized in Table 4. Specifically, in each repetition, human texts are independently split into $\mathcal{D}^0_{\text{pilot}}$ (20%), $\mathcal{D}^0_{\text{dev}}$ (20%, divided into two equal halves for cross-fitting the trust radius), $\mathcal{D}_{\text{cal}}$ (30%), and the null candidates of $\mathcal{D}_{\text{test}}$ (30%). Machine texts are similarly partitioned into $\mathcal{D}^1_{\text{pilot}}$ (20%), $\mathcal{D}^1_{\text{dev}}$ (20%), and the non-null candidates of $\mathcal{D}_{\text{test}}$ (30%). All reported empirical FDR and selection power values are Monte Carlo averages over 1000 independent repetitions of this partition, with the baseline and our method evaluated on the exact same role splits in each repetition.

Table 4: Role sizes for the twelve in-domain candidate pools. Superscripts 0 and 1 denote human and machine-generated text. m is the candidate-pool size and π_1 the fraction of it that is machine-generated. Development counts are the sum of the two cross-fitting folds.

Domain	**Generator**	n^0_{pilot}	n^1_{pilot}	n^0_{dev}	n^1_{dev}	n_{cal}	m	π_1
ChangeMyView	LLaMA-7B	1812	156	1812	156	2719	2953	0.079
	GPT-3.5 Turbo	1812	482	1812	482	2719	3442	0.210
	Davinci-003	1812	480	1812	480	2719	3440	0.210
HellaSwag	LLaMA-7B	1942	193	1942	192	2913	3201	0.090
	GPT-3.5 Turbo	1942	599	1942	599	2913	3810	0.236
	Davinci-003	1942	501	1942	500	2913	3663	0.205
WritingPrompts	LLaMA-7B	2518	192	2518	192	3776	4064	0.071
	GPT-3.5 Turbo	2518	599	2518	599	3776	4674	0.192
	Davinci-003	2518	599	2518	598	3776	4674	0.192
XSum	LLaMA-7B	2253	199	2254	200	3379	3677	0.081
	GPT-3.5 Turbo	2253	600	2254	600	3379	4279	0.210
	Davinci-003	2253	599	2254	600	3379	4278	0.210

Fixed constants and the selected hyperparameter. All algorithmic parameters except the trust radius are fixed protocol constants across every reported run and were never tuned per setting: the feature map projects GPT-2 representations by PCA to $d = 32$ dimensions (with the encoder never updated), the smoothing bandwidth is $h = 0.25$, the curvature ridge is $\gamma = 10^{-3}$, and the pilot null-proportion surrogate in Equation (32) uses $\lambda_{\text{pilot}} = 0.5$. For the final Storey–BH rule, the threshold is set to $\lambda = K/(n_{\text{cal}} + 1)$ with the integer K placing λ closest to 0.5. The radius r is the single selected hyperparameter, chosen separately for each target level by development cross-fitting from the pre-registered grid $\{0.001, 0.0025, 0.005, 0.01, 0.02, 0.05, 0.10, 0.20, 0.40\}$. The radius is chosen using development data only, as the pooled equivariance Equation (29) requires.

Computational cost. Runs are CPU-only, with no GPU used at selection time; detector scores are computed once and cached, so the cost reported here is the cost of the selection experiments alone. On 32 CPU cores, one (detection score, candidate pool) combination at 1000 repetitions and four target levels takes a median of 9.3 minutes of wall-clock time, with a range of 8.7 to 9.7 minutes over 160 measured runs. The full four-domain grid of 108 combinations is therefore roughly 19 hours of 32-core wall-clock time. Memory use fits in 16 GB.

D Additional Experiments

D.1 Full results on the primary datasets

We report selection power and the empirical FDR for every cell of the primary grid in Section 4.2. Tables 6 to 9 give selection power and Tables 10 to 13 give the empirical FDR, one table per dataset, for every generator, detection score, and target over 1000 repetitions. Table 5 reports the same selection power with the nine detection scores compressed to their median. The results show that our method attains higher selection power than the baseline on these pools.

Table 5: Selection power on the four primary domains, resolved by domain, generator, and target level, with the nine detection scores compressed to their median. **Bold** marks the larger value in each pair.

Domain	**Generator**	$q = 0.05$		$q = 0.10$		$q = 0.20$		$q = 0.30$	
		Base	Ours	Base	Ours	Base	Ours	Base	Ours
ChangeMyView	LLaMA-7B	0.848	**0.850**	0.902	**0.907**	0.927	**0.938**	0.941	**0.949**
	GPT-3.5 Turbo	0.105	**0.314**	0.286	**0.481**	0.528	**0.745**	0.635	**0.866**
	Davinci-003	0.148	**0.467**	0.271	**0.580**	0.436	**0.749**	0.760	**0.901**
HellaSwag	LLaMA-7B	0.694	**0.788**	0.741	**0.831**	0.773	**0.866**	0.807	**0.888**
	GPT-3.5 Turbo	0.013	**0.043**	0.062	**0.291**	0.235	**0.487**	0.338	**0.572**
	Davinci-003	0.029	**0.035**	0.060	**0.215**	0.273	**0.501**	0.465	**0.601**
WritingPrompts	LLaMA-7B	0.811	**0.839**	0.874	**0.881**	0.905	**0.923**	0.917	**0.935**
	GPT-3.5 Turbo	0.355	**0.631**	0.535	**0.703**	0.645	**0.761**	0.703	**0.879**
	Davinci-003	0.482	**0.561**	0.594	**0.695**	0.712	**0.805**	0.785	**0.918**
XSum	LLaMA-7B	0.007	**0.008**	0.049	**0.099**	0.255	**0.484**	0.518	**0.586**
	GPT-3.5 Turbo	0.333	**0.431**	0.420	**0.590**	0.599	**0.713**	0.671	**0.797**
	Davinci-003	0.110	**0.333**	0.319	**0.589**	0.560	**0.741**	0.671	**0.812**

D.2 Evaluation on more datasets

In this section, we apply the identical setting to six further datasets, Yelp, ELI5, SQuAD, ROCStories, SciXGen, and TLDR, with LLaMA-7B as the generator and the same nine detection scores, over 1000 repetitions at each target level. Table 14 reports selection power for every dataset, detection score, and target level. The results show that our method raises selection power over the baseline on all six datasets.

Table 6: Selection power on the ChangeMyView pools of the primary grid, resolved by generator, detection score, and target level. Each entry is the mean over 1000 repetitions; nothing is averaged across generators or targets. **Bold** marks the larger value in each pair.

Generator	**Detection score**	$q = 0.05$		$q = 0.10$		$q = 0.20$		$q = 0.30$	
		Base	Ours	Base	Ours	Base	Ours	Base	Ours
LLaMA-7B	Log-Likelihood	0.914	**0.921**	0.947	**0.948**	0.958	**0.960**	0.966	**0.967**
	Rank	0.386	**0.577**	0.532	**0.697**	0.689	**0.796**	0.776	**0.849**
	Log-Rank	0.925	**0.928**	0.947	**0.948**	0.958	**0.960**	0.969	0.969
	Entropy	0.559	**0.586**	0.658	**0.684**	0.729	**0.754**	0.779	**0.797**
	DetectLLM-LRR	0.848	**0.850**	0.908	**0.907**	0.938	**0.941**	0.955	**0.957**
	Fast-DetectGPT	0.936	**0.938**	0.943	**0.947**	0.953	**0.956**	0.957	**0.962**
	AdaDetectGPT	0.875	**0.892**	0.902	**0.917**	0.927	**0.938**	0.941	**0.949**
	RoBERTa-base	0.820	**0.825**	0.864	**0.868**	0.901	**0.903**	0.924	**0.925**
	RADAR	0.000	0.000	0.000	0.000	0.000	**0.001**	0.000	**0.005**
GPT-3.5 Turbo	Log-Likelihood	0.607	**0.798**	0.736	**0.873**	0.830	**0.921**	0.873	**0.947**
	Rank	0.105	**0.586**	0.286	**0.732**	0.528	**0.796**	0.635	**0.844**
	Log-Rank	0.551	**0.775**	0.700	**0.849**	0.808	**0.911**	0.857	**0.941**
	Entropy	0.002	**0.011**	0.019	**0.149**	0.181	**0.559**	0.334	**0.682**
	DetectLLM-LRR	0.205	**0.314**	0.442	**0.664**	0.620	**0.796**	0.722	**0.866**
	Fast-DetectGPT	0.024	**0.092**	0.138	**0.450**	0.393	**0.745**	0.618	**0.851**
	AdaDetectGPT	0.000	**0.001**	0.002	**0.036**	0.030	**0.350**	0.209	**0.668**
	RoBERTa-base	0.223	**0.439**	0.369	**0.481**	0.554	**0.603**	0.708	**0.880**
	RADAR	0.000	**0.012**	0.005	**0.162**	0.030	**0.663**	0.137	**0.943**
Davinci-003	Log-Likelihood	0.634	**0.815**	0.761	**0.895**	0.863	**0.941**	0.907	**0.960**
	Rank	0.148	**0.484**	0.271	**0.593**	0.436	**0.662**	0.515	**0.707**
	Log-Rank	0.605	**0.802**	0.737	**0.882**	0.849	**0.935**	0.895	**0.958**
	Entropy	0.000	**0.001**	0.007	**0.044**	0.068	**0.350**	0.179	**0.557**
	DetectLLM-LRR	0.344	**0.467**	0.544	**0.738**	0.696	**0.847**	0.784	**0.901**
	Fast-DetectGPT	0.009	**0.043**	0.072	**0.353**	0.332	**0.766**	0.626	**0.884**
	AdaDetectGPT	0.000	0.000	0.002	**0.019**	0.016	**0.281**	0.127	**0.651**
	RoBERTa-base	0.286	**0.514**	0.489	**0.580**	0.722	**0.749**	0.843	**0.914**
	RADAR	0.000	**0.001**	0.000	**0.026**	0.020	**0.295**	0.760	**0.951**

Table 7: Selection power on the HellaSwag pools of the primary grid, resolved by generator, detection score, and target level. Each entry is one combination over 1000 repetitions; nothing is averaged across generators or targets. **Bold** marks the larger value in each pair.

Generator	**Detection score**	$q = 0.05$		$q = 0.10$		$q = 0.20$		$q = 0.30$	
		Base	Ours	Base	Ours	Base	Ours	Base	Ours
LLaMA-7B	Log-Likelihood	0.776	**0.861**	0.819	**0.888**	0.859	**0.911**	0.886	**0.924**
	Rank	0.000	**0.006**	0.002	**0.073**	0.044	**0.405**	0.199	**0.737**
	Log-Rank	0.728	**0.847**	0.789	**0.883**	0.849	**0.913**	0.880	**0.926**
	Entropy	0.672	**0.788**	0.726	**0.831**	0.771	**0.866**	0.807	**0.888**
	DetectLLM-LRR	0.010	**0.011**	0.145	**0.260**	0.631	**0.774**	0.753	**0.849**
	Fast-DetectGPT	0.877	**0.887**	0.892	**0.901**	0.905	**0.917**	0.913	**0.928**
	AdaDetectGPT	0.859	**0.878**	0.882	**0.895**	0.902	**0.911**	0.909	**0.923**
	RoBERTa-base	0.694	**0.752**	0.741	**0.784**	0.773	**0.813**	0.792	**0.827**
	RADAR	0.449	0.449	0.562	**0.569**	0.665	**0.707**	0.705	**0.764**
GPT-3.5 Turbo	Log-Likelihood	0.106	**0.315**	0.228	**0.463**	0.412	**0.576**	0.530	**0.653**
	Rank	0.000	0.000	0.000	**0.002**	0.000	**0.029**	0.000	**0.123**
	Log-Rank	0.004	**0.043**	0.062	**0.291**	0.349	**0.558**	0.506	**0.640**
	Entropy	0.013	**0.071**	0.080	**0.291**	0.235	**0.487**	0.338	**0.572**
	DetectLLM-LRR	0.000	0.000	0.000	0.000	0.000	0.000	0.003	**0.090**
	Fast-DetectGPT	0.381	**0.446**	0.439	**0.502**	0.498	**0.569**	0.550	**0.632**
	AdaDetectGPT	0.290	**0.390**	0.380	**0.465**	0.458	**0.544**	0.522	**0.617**
	RoBERTa-base	0.025	0.025	0.048	**0.068**	0.079	**0.121**	0.110	**0.188**
	RADAR	0.013	0.013	0.043	**0.044**	0.141	**0.194**	0.214	**0.403**
Davinci-003	Log-Likelihood	0.071	**0.156**	0.161	**0.362**	0.360	**0.543**	0.509	**0.649**
	Rank	0.000	0.000	0.000	**0.006**	0.001	**0.050**	0.005	**0.189**
	Log-Rank	0.008	**0.032**	0.060	**0.215**	0.273	**0.512**	0.483	**0.632**
	Entropy	0.309	**0.467**	0.373	**0.577**	0.458	**0.670**	0.551	**0.730**
	DetectLLM-LRR	0.000	0.000	0.000	0.000	0.001	**0.003**	0.026	**0.137**
	Fast-DetectGPT	0.255	**0.323**	0.352	**0.443**	0.444	**0.557**	0.542	**0.660**
	AdaDetectGPT	0.136	**0.202**	0.243	**0.361**	0.371	**0.501**	0.465	**0.601**
	RoBERTa-base	0.029	**0.035**	0.055	**0.103**	0.095	**0.185**	0.137	**0.264**
	RADAR	0.002	0.002	0.012	0.012	0.035	**0.036**	0.063	**0.102**

Table 8: Selection power on the WritingPrompts pools of the primary grid, resolved by generator, detection score, and target level. Each entry is one combination over 1000 repetitions; nothing is averaged across generators or targets. **Bold** marks the larger value in each pair.

Generator	**Detection score**	$q = 0.05$		$q = 0.10$		$q = 0.20$		$q = 0.30$	
		Base	Ours	Base	Ours	Base	Ours	Base	Ours
LLaMA-7B	Log-Likelihood	0.840	**0.849**	0.912	**0.918**	0.936	**0.946**	0.954	**0.958**
	Rank	0.383	**0.535**	0.550	**0.701**	0.693	**0.805**	0.767	**0.855**
	Log-Rank	0.862	**0.866**	0.915	**0.920**	0.940	**0.946**	0.952	**0.956**
	Entropy	0.323	**0.333**	0.564	**0.600**	0.699	**0.747**	0.767	**0.803**
	DetectLLM-LRR	0.832	**0.839**	0.881	**0.893**	0.913	**0.923**	0.928	**0.937**
	Fast-DetectGPT	0.871	**0.891**	0.883	**0.909**	0.905	**0.925**	0.917	**0.935**
	AdaDetectGPT	0.811	**0.844**	0.831	**0.868**	0.856	**0.896**	0.875	**0.911**
	RoBERTa-base	0.800	**0.812**	0.874	**0.881**	0.922	**0.923**	0.947	**0.948**
	RADAR	0.000	0.000	0.002	**0.003**	0.012	**0.035**	0.027	**0.098**
GPT-3.5 Turbo	Log-Likelihood	0.659	**0.795**	0.765	**0.849**	0.817	**0.887**	0.847	**0.908**
	Rank	0.355	**0.717**	0.537	**0.779**	0.675	**0.829**	0.738	**0.856**
	Log-Rank	0.649	**0.795**	0.755	**0.847**	0.812	**0.884**	0.846	**0.904**
	Entropy	0.000	0.000	0.001	**0.020**	0.264	**0.593**	0.531	**0.762**
	DetectLLM-LRR	0.507	**0.694**	0.655	**0.796**	0.752	**0.853**	0.801	**0.879**
	Fast-DetectGPT	0.427	**0.631**	0.535	**0.703**	0.645	**0.761**	0.703	**0.800**
	AdaDetectGPT	0.171	**0.431**	0.269	**0.590**	0.468	**0.683**	0.586	**0.742**
	RoBERTa-base	0.344	**0.451**	0.457	**0.496**	0.589	**0.638**	0.686	**0.932**
	RADAR	0.187	**0.274**	0.221	**0.364**	0.361	**0.506**	0.616	**0.883**
Davinci-003	Log-Likelihood	0.760	**0.857**	0.838	**0.891**	0.879	**0.915**	0.900	**0.928**
	Rank	0.519	**0.735**	0.614	**0.786**	0.713	**0.829**	0.763	**0.856**
	Log-Rank	0.763	**0.863**	0.837	**0.893**	0.877	**0.917**	0.901	**0.930**
	Entropy	0.000	0.000	0.000	**0.007**	0.124	**0.458**	0.490	**0.759**
	DetectLLM-LRR	0.701	**0.821**	0.789	**0.876**	0.849	**0.906**	0.883	**0.923**
	Fast-DetectGPT	0.266	**0.549**	0.419	**0.695**	0.613	**0.798**	0.731	**0.850**
	AdaDetectGPT	0.070	**0.278**	0.171	**0.558**	0.375	**0.703**	0.564	**0.782**
	RoBERTa-base	0.482	**0.561**	0.594	**0.614**	0.712	**0.765**	0.785	**0.918**
	RADAR	0.080	**0.275**	0.166	**0.487**	0.647	**0.805**	0.882	**0.969**

Table 9: Selection power on the XSum pools of the primary grid, resolved by generator, detection score, and target level. Each entry is one combination over 1000 repetitions; nothing is averaged across generators or targets. **Bold** marks the larger value in each pair.

Generator	**Detection score**	$q = 0.05$		$q = 0.10$		$q = 0.20$		$q = 0.30$	
		Base	Ours	Base	Ours	Base	Ours	Base	Ours
LLaMA-7B	Log-Likelihood	0.001	**0.002**	0.016	**0.018**	0.044	**0.112**	0.204	**0.435**
	Rank	0.002	0.002	0.016	**0.028**	0.083	**0.117**	0.166	**0.216**
	Log-Rank	0.007	**0.008**	0.049	**0.099**	0.255	**0.484**	0.518	**0.633**
	Entropy	0.000	0.000	0.003	0.003	0.015	**0.016**	0.024	**0.028**
	DetectLLM-LRR	0.059	**0.065**	0.294	**0.351**	0.460	**0.508**	0.561	**0.586**
	Fast-DetectGPT	0.512	**0.532**	0.578	**0.605**	0.643	**0.674**	0.685	**0.716**
	AdaDetectGPT	0.451	**0.471**	0.519	**0.536**	0.562	**0.593**	0.602	**0.647**
	RoBERTa-base	0.504	**0.509**	0.613	0.613	0.718	0.718	0.778	0.778
	RADAR	0.000	0.000	0.000	0.000	0.001	0.001	0.003	0.003
GPT-3.5 Turbo	Log-Likelihood	0.000	**0.001**	0.007	**0.074**	0.576	**0.744**	0.671	**0.790**
	Rank	0.509	**0.813**	0.581	**0.846**	0.688	**0.884**	0.751	**0.910**
	Log-Rank	0.087	**0.280**	0.420	**0.642**	0.618	**0.759**	0.686	**0.797**
	Entropy	0.333	**0.336**	0.334	**0.371**	0.362	**0.713**	0.478	**0.847**
	DetectLLM-LRR	0.422	**0.576**	0.517	**0.669**	0.624	**0.746**	0.678	**0.783**
	Fast-DetectGPT	0.461	**0.587**	0.541	**0.634**	0.599	**0.655**	0.624	**0.664**
	AdaDetectGPT	0.224	**0.431**	0.358	**0.560**	0.486	**0.620**	0.557	**0.645**
	RoBERTa-base	0.467	**0.543**	0.564	**0.590**	0.686	**0.706**	0.770	**0.884**
	RADAR	0.114	**0.221**	0.146	**0.444**	0.198	**0.585**	0.561	**0.899**
Davinci-003	Log-Likelihood	0.000	0.000	0.029	**0.141**	0.695	**0.838**	0.788	**0.878**
	Rank	0.212	**0.491**	0.319	**0.589**	0.437	**0.675**	0.521	**0.732**
	Log-Rank	0.110	**0.333**	0.521	**0.733**	0.732	**0.846**	0.801	**0.880**
	Entropy	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.001**
	DetectLLM-LRR	0.492	**0.624**	0.598	**0.730**	0.725	**0.811**	0.784	**0.854**
	Fast-DetectGPT	0.204	**0.405**	0.401	**0.626**	0.560	**0.741**	0.671	**0.812**
	AdaDetectGPT	0.056	**0.174**	0.199	**0.473**	0.371	**0.641**	0.531	**0.739**
	RoBERTa-base	0.540	**0.585**	0.635	**0.648**	0.739	**0.751**	0.815	**0.850**
	RADAR	0.000	0.000	0.000	**0.006**	0.001	**0.075**	0.005	**0.216**

Table 10: Empirical FDR on the ChangeMyView pools of the primary grid, resolved by generator, detection score, and target level. Each entry is one combination over 1000 repetitions; nothing is averaged across generators or targets.

Generator	**Detection score**	$q = 0.05$		$q = 0.10$		$q = 0.20$		$q = 0.30$	
		Base	Ours	Base	Ours	Base	Ours	Base	Ours
LLaMA-7B	Log-Likelihood	0.049	0.049	0.100	0.100	0.202	0.201	0.299	0.300
	Rank	0.047	0.048	0.099	0.098	0.201	0.200	0.300	0.299
	Log-Rank	0.048	0.049	0.100	0.100	0.200	0.200	0.300	0.300
	Entropy	0.048	0.048	0.100	0.100	0.200	0.200	0.300	0.300
	DetectLLM-LRR	0.049	0.050	0.100	0.099	0.200	0.200	0.300	0.300
	Fast-DetectGPT	0.048	0.047	0.098	0.098	0.200	0.199	0.299	0.300
	AdaDetectGPT	0.047	0.048	0.098	0.098	0.198	0.199	0.299	0.300
	RoBERTa-base	0.048	0.047	0.098	0.098	0.198	0.198	0.298	0.299
	RADAR	0.000	0.000	0.001	0.001	0.015	0.020	0.073	0.095
GPT-3.5 Turbo	Log-Likelihood	0.048	0.048	0.097	0.098	0.196	0.198	0.295	0.298
	Rank	0.044	0.048	0.094	0.097	0.190	0.194	0.286	0.292
	Log-Rank	0.048	0.048	0.097	0.098	0.196	0.197	0.294	0.298
	Entropy	0.007	0.009	0.049	0.066	0.181	0.191	0.278	0.289
	DetectLLM-LRR	0.046	0.047	0.096	0.097	0.193	0.196	0.289	0.295
	Fast-DetectGPT	0.032	0.035	0.093	0.097	0.195	0.196	0.292	0.296
	AdaDetectGPT	0.001	0.002	0.030	0.038	0.157	0.183	0.279	0.291
	RoBERTa-base	0.047	0.048	0.097	0.099	0.197	0.198	0.296	0.299
	RADAR	0.004	0.005	0.044	0.057	0.180	0.192	0.295	0.299
Davinci-003	Log-Likelihood	0.048	0.049	0.099	0.099	0.198	0.200	0.298	0.299
	Rank	0.043	0.044	0.088	0.091	0.179	0.184	0.269	0.276
	Log-Rank	0.048	0.049	0.099	0.099	0.198	0.199	0.298	0.300
	Entropy	0.003	0.003	0.040	0.050	0.158	0.182	0.266	0.286
	DetectLLM-LRR	0.047	0.048	0.098	0.099	0.195	0.198	0.293	0.297
	Fast-DetectGPT	0.019	0.023	0.090	0.096	0.197	0.199	0.296	0.300
	AdaDetectGPT	0.001	0.001	0.027	0.035	0.145	0.169	0.284	0.296
	RoBERTa-base	0.047	0.048	0.099	0.099	0.200	0.199	0.299	0.299
	RADAR	0.000	0.000	0.001	0.005	0.046	0.095	0.298	0.300

Table 11: Empirical FDR on the HellaSwag pools of the primary grid, resolved by generator, detection score, and target level. Each entry is one combination over 1000 repetitions; nothing is averaged across generators or targets.

Generator	**Detection score**	$q = 0.05$		$q = 0.10$		$q = 0.20$		$q = 0.30$	
		Base	Ours	Base	Ours	Base	Ours	Base	Ours
LLaMA-7B	Log-Likelihood	0.048	0.049	0.099	0.099	0.199	0.199	0.300	0.300
	Rank	0.000	0.001	0.011	0.020	0.128	0.161	0.282	0.295
	Log-Rank	0.049	0.048	0.099	0.098	0.199	0.200	0.300	0.300
	Entropy	0.046	0.047	0.096	0.098	0.197	0.198	0.296	0.298
	DetectLLM-LRR	0.010	0.011	0.068	0.078	0.200	0.200	0.298	0.300
	Fast-DetectGPT	0.048	0.049	0.098	0.098	0.199	0.200	0.299	0.299
	AdaDetectGPT	0.047	0.048	0.097	0.098	0.199	0.199	0.299	0.300
	RoBERTa-base	0.045	0.047	0.096	0.097	0.195	0.196	0.293	0.294
	RADAR	0.047	0.047	0.097	0.098	0.197	0.199	0.297	0.300
GPT-3.5 Turbo	Log-Likelihood	0.042	0.044	0.091	0.093	0.186	0.188	0.279	0.284
	Rank	0.000	0.000	0.001	0.002	0.017	0.030	0.065	0.115
	Log-Rank	0.010	0.015	0.074	0.086	0.185	0.187	0.278	0.282
	Entropy	0.025	0.031	0.086	0.089	0.181	0.184	0.273	0.278
	DetectLLM-LRR	0.000	0.000	0.001	0.001	0.015	0.015	0.086	0.128
	Fast-DetectGPT	0.045	0.045	0.091	0.093	0.182	0.187	0.274	0.281
	AdaDetectGPT	0.045	0.045	0.091	0.092	0.182	0.186	0.275	0.281
	RoBERTa-base	0.030	0.029	0.072	0.075	0.155	0.165	0.239	0.268
	RADAR	0.026	0.026	0.077	0.077	0.170	0.175	0.257	0.276
Davinci-003	Log-Likelihood	0.042	0.045	0.093	0.094	0.190	0.191	0.285	0.287
	Rank	0.000	0.000	0.001	0.004	0.029	0.052	0.121	0.188
	Log-Rank	0.015	0.021	0.082	0.090	0.189	0.190	0.284	0.286
	Entropy	0.046	0.047	0.094	0.096	0.190	0.193	0.286	0.289
	DetectLLM-LRR	0.000	0.000	0.003	0.003	0.070	0.072	0.183	0.226
	Fast-DetectGPT	0.047	0.047	0.096	0.097	0.192	0.194	0.288	0.292
	AdaDetectGPT	0.044	0.046	0.094	0.094	0.192	0.192	0.286	0.289
	RoBERTa-base	0.032	0.034	0.078	0.082	0.165	0.171	0.253	0.262
	RADAR	0.010	0.010	0.057	0.057	0.161	0.161	0.254	0.260

Table 12: Empirical FDR on the WritingPrompts pools of the primary grid, resolved by generator, detection score, and target level. Each entry is one combination over 1000 repetitions; nothing is averaged across generators or targets.

Generator	**Detection score**	$q = 0.05$		$q = 0.10$		$q = 0.20$		$q = 0.30$	
		Base	Ours	Base	Ours	Base	Ours	Base	Ours
LLaMA-7B	Log-Likelihood	0.048	0.048	0.098	0.099	0.198	0.199	0.299	0.299
	Rank	0.046	0.047	0.097	0.097	0.198	0.198	0.298	0.297
	Log-Rank	0.048	0.049	0.098	0.099	0.199	0.200	0.299	0.299
	Entropy	0.046	0.047	0.097	0.098	0.195	0.196	0.295	0.296
	DetectLLM-LRR	0.048	0.048	0.098	0.097	0.198	0.197	0.299	0.298
	Fast-DetectGPT	0.048	0.048	0.098	0.099	0.199	0.199	0.298	0.300
	AdaDetectGPT	0.048	0.049	0.099	0.099	0.198	0.199	0.297	0.298
	RoBERTa-base	0.048	0.048	0.099	0.100	0.198	0.198	0.300	0.300
	RADAR	0.000	0.000	0.022	0.022	0.136	0.151	0.261	0.282
GPT-3.5 Turbo	Log-Likelihood	0.048	0.049	0.097	0.098	0.193	0.196	0.290	0.295
	Rank	0.047	0.048	0.096	0.097	0.192	0.194	0.287	0.290
	Log-Rank	0.048	0.049	0.097	0.097	0.193	0.195	0.290	0.295
	Entropy	0.000	0.000	0.001	0.005	0.168	0.191	0.287	0.292
	DetectLLM-LRR	0.048	0.048	0.096	0.097	0.193	0.195	0.288	0.293
	Fast-DetectGPT	0.047	0.047	0.092	0.094	0.186	0.190	0.281	0.288
	AdaDetectGPT	0.046	0.047	0.092	0.093	0.185	0.188	0.278	0.283
	RoBERTa-base	0.048	0.048	0.096	0.098	0.193	0.197	0.290	0.300
	RADAR	0.046	0.048	0.096	0.099	0.199	0.199	0.299	0.300
Davinci-003	Log-Likelihood	0.048	0.049	0.098	0.098	0.196	0.198	0.294	0.296
	Rank	0.048	0.048	0.096	0.097	0.192	0.194	0.289	0.292
	Log-Rank	0.048	0.049	0.098	0.098	0.196	0.198	0.294	0.296
	Entropy	0.000	0.000	0.000	0.002	0.111	0.164	0.286	0.289
	DetectLLM-LRR	0.048	0.048	0.097	0.098	0.196	0.197	0.294	0.296
	Fast-DetectGPT	0.048	0.049	0.097	0.097	0.195	0.196	0.292	0.295
	AdaDetectGPT	0.042	0.046	0.094	0.097	0.193	0.194	0.290	0.291
	RoBERTa-base	0.048	0.048	0.097	0.098	0.196	0.198	0.294	0.299
	RADAR	0.047	0.050	0.100	0.100	0.200	0.200	0.300	0.300

Table 13: Empirical FDR on the XSum pools of the primary grid, resolved by generator, detection score, and target level. Each entry is one combination over 1000 repetitions; nothing is averaged across generators or targets.

Generator	**Detection score**	$q = 0.05$		$q = 0.10$		$q = 0.20$		$q = 0.30$	
		Base	Ours	Base	Ours	Base	Ours	Base	Ours
LLaMA-7B	Log-Likelihood	0.006	0.006	0.057	0.059	0.170	0.179	0.283	0.295
	Rank	0.006	0.007	0.052	0.062	0.182	0.185	0.286	0.289
	Log-Rank	0.014	0.014	0.075	0.080	0.191	0.196	0.297	0.297
	Entropy	0.000	0.000	0.030	0.030	0.143	0.143	0.254	0.254
	DetectLLM-LRR	0.026	0.027	0.093	0.095	0.194	0.197	0.292	0.293
	Fast-DetectGPT	0.048	0.048	0.099	0.098	0.198	0.196	0.295	0.296
	AdaDetectGPT	0.047	0.047	0.098	0.098	0.197	0.195	0.295	0.294
	RoBERTa-base	0.047	0.046	0.096	0.096	0.199	0.199	0.299	0.299
	RADAR	0.000	0.000	0.002	0.002	0.069	0.068	0.169	0.169
GPT-3.5 Turbo	Log-Likelihood	0.000	0.000	0.004	0.014	0.183	0.187	0.275	0.283
	Rank	0.047	0.049	0.096	0.098	0.193	0.197	0.291	0.296
	Log-Rank	0.028	0.037	0.090	0.092	0.183	0.188	0.276	0.284
	Entropy	0.047	0.047	0.094	0.096	0.191	0.197	0.289	0.297
	DetectLLM-LRR	0.045	0.046	0.092	0.093	0.185	0.189	0.278	0.286
	Fast-DetectGPT	0.042	0.042	0.085	0.085	0.171	0.171	0.257	0.258
	AdaDetectGPT	0.041	0.042	0.085	0.084	0.169	0.169	0.254	0.254
	RoBERTa-base	0.047	0.048	0.096	0.097	0.195	0.196	0.293	0.298
	RADAR	0.047	0.047	0.096	0.099	0.197	0.200	0.300	0.300
Davinci-003	Log-Likelihood	0.000	0.000	0.011	0.025	0.194	0.195	0.291	0.294
	Rank	0.045	0.047	0.092	0.095	0.184	0.191	0.278	0.286
	Log-Rank	0.033	0.041	0.096	0.096	0.193	0.195	0.290	0.293
	Entropy	0.000	0.000	0.000	0.000	0.019	0.019	0.070	0.073
	DetectLLM-LRR	0.047	0.048	0.095	0.097	0.191	0.195	0.287	0.293
	Fast-DetectGPT	0.046	0.047	0.096	0.096	0.192	0.195	0.289	0.293
	AdaDetectGPT	0.033	0.039	0.093	0.096	0.189	0.193	0.286	0.289
	RoBERTa-base	0.048	0.048	0.097	0.098	0.194	0.195	0.292	0.296
	RADAR	0.000	0.000	0.009	0.011	0.087	0.110	0.200	0.247

Table 14: Selection power on the six further MAGE datasets with LLaMA-7B, resolved by dataset, detection score, and target level. Each entry is one combination over 1000 repetitions; nothing is averaged across datasets, detection scores, or targets. **Bold** marks the larger value in each pair.

Dataset	Detection score	$q = 0.05$		$q = 0.10$		$q = 0.20$		$q = 0.30$	
		Base	Ours	Base	Ours	Base	Ours	Base	Ours
ELI5	Log-Likelihood	0.648	**0.691**	0.762	**0.791**	0.823	**0.842**	0.846	**0.866**
	Rank	0.006	**0.024**	0.056	**0.224**	0.237	**0.503**	0.379	**0.600**
	Log-Rank	0.684	**0.721**	0.769	**0.796**	0.829	**0.844**	0.848	**0.866**
	Entropy	0.235	**0.245**	0.356	**0.392**	0.501	**0.551**	0.566	**0.623**
	DetectLLM-LRR	0.448	**0.482**	0.582	**0.641**	0.731	**0.764**	0.784	**0.815**
	Fast-DetectGPT	0.775	**0.789**	0.805	**0.816**	0.834	**0.847**	0.855	**0.869**
	AdaDetectGPT	0.685	**0.716**	0.746	**0.768**	0.787	**0.810**	0.812	**0.835**
	RoBERTa-base	0.486	**0.584**	0.694	**0.712**	0.749	**0.767**	0.779	**0.794**
	RADAR	0.000	0.000	0.000	0.000	0.000	**0.001**	0.000	**0.002**
ROCStories	Log-Likelihood	0.786	**0.820**	0.815	**0.855**	0.857	**0.887**	0.879	**0.905**
	Rank	0.178	**0.482**	0.234	**0.607**	0.309	**0.713**	0.383	**0.763**
	Log-Rank	0.745	**0.803**	0.791	**0.846**	0.845	**0.880**	0.876	**0.900**
	Entropy	0.711	**0.773**	0.751	**0.816**	0.795	**0.850**	0.832	**0.869**
	DetectLLM-LRR	0.241	**0.289**	0.481	**0.602**	0.664	**0.764**	0.744	**0.821**
	Fast-DetectGPT	0.900	**0.904**	0.912	**0.917**	0.925	**0.928**	0.931	**0.936**
	AdaDetectGPT	0.883	**0.890**	0.893	**0.904**	0.916	**0.922**	0.930	**0.934**
	RoBERTa-base	0.379	**0.747**	0.839	**0.847**	0.865	**0.878**	0.882	**0.893**
	RADAR	0.661	0.661	0.730	**0.740**	0.797	**0.825**	0.831	**0.858**
SciXGen	Log-Likelihood	0.514	**0.527**	0.588	**0.614**	0.668	**0.694**	0.709	**0.756**
	Rank	0.105	**0.141**	0.180	**0.263**	0.266	**0.389**	0.355	**0.473**
	Log-Rank	0.528	**0.545**	0.613	**0.630**	0.679	**0.709**	0.722	**0.769**
	Entropy	0.298	**0.305**	0.356	**0.372**	0.406	**0.424**	0.437	**0.468**
	DetectLLM-LRR	0.333	**0.337**	0.458	**0.471**	0.578	**0.604**	0.638	**0.683**
	Fast-DetectGPT	0.644	**0.650**	0.691	**0.704**	0.757	**0.771**	0.801	**0.814**
	AdaDetectGPT	0.596	**0.604**	0.659	**0.668**	0.715	**0.734**	0.760	**0.783**
	RoBERTa-base	0.194	**0.237**	0.339	**0.341**	0.419	**0.432**	0.501	**0.506**
	RADAR	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001
SQuAD	Log-Likelihood	0.257	**0.308**	0.426	**0.523**	0.582	**0.681**	0.669	**0.760**
	Rank	0.000	**0.002**	0.012	**0.033**	0.045	**0.173**	0.103	**0.326**
	Log-Rank	0.318	**0.403**	0.460	**0.574**	0.615	**0.705**	0.707	**0.777**
	Entropy	0.105	**0.109**	0.151	**0.171**	0.205	**0.261**	0.255	**0.336**
	DetectLLM-LRR	0.089	**0.097**	0.247	**0.286**	0.451	**0.506**	0.560	**0.625**
	Fast-DetectGPT	0.876	**0.877**	0.888	**0.889**	0.903	**0.905**	0.913	**0.917**
	AdaDetectGPT	0.849	**0.851**	0.869	0.869	0.888	**0.891**	0.902	**0.906**
	RoBERTa-base	0.002	**0.023**	0.274	**0.354**	0.621	**0.645**	0.683	**0.703**
	RADAR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
TLDR	Log-Likelihood	0.651	**0.668**	0.700	**0.732**	0.791	**0.811**	0.840	**0.857**
	Rank	0.003	**0.006**	0.026	**0.057**	0.112	**0.203**	0.215	**0.312**
	Log-Rank	0.646	**0.665**	0.710	**0.733**	0.787	**0.815**	0.848	**0.864**
	Entropy	0.203	**0.221**	0.271	**0.330**	0.346	**0.451**	0.425	**0.523**
	DetectLLM-LRR	0.301	**0.308**	0.487	**0.504**	0.640	**0.671**	0.737	**0.758**
	Fast-DetectGPT	0.844	**0.848**	0.861	**0.872**	0.896	**0.901**	0.917	**0.918**
	AdaDetectGPT	0.845	**0.846**	0.863	**0.865**	0.884	**0.885**	0.896	**0.903**
	RoBERTa-base	0.001	**0.010**	0.092	**0.143**	0.703	**0.710**	0.750	**0.752**
	RADAR	0.000	0.000	0.000	0.000	0.000	**0.009**	0.001	**0.038**
Yelp	Log-Likelihood	0.444	**0.527**	0.521	**0.597**	0.599	**0.660**	0.655	**0.697**
	Rank	0.000	**0.001**	0.001	**0.008**	0.008	**0.083**	0.024	**0.190**
	Log-Rank	0.382	**0.513**	0.501	**0.595**	0.598	**0.659**	0.655	**0.698**
	Entropy	0.221	**0.234**	0.273	**0.306**	0.318	**0.359**	0.346	**0.399**
	DetectLLM-LRR	0.002	0.002	0.050	**0.065**	0.344	**0.462**	0.485	**0.573**
	Fast-DetectGPT	0.651	**0.668**	0.683	**0.699**	0.718	**0.732**	0.739	**0.756**
	AdaDetectGPT	0.594	**0.623**	0.643	**0.665**	0.680	**0.696**	0.701	**0.720**
	RoBERTa-base	0.007	**0.026**	0.143	**0.247**	0.618	**0.622**	0.654	**0.658**
	RADAR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.002**