REFORMULATING DATA-ADAPTIVE LLM-GENERATED TEXT DETECTION: A CLASS-CONDITIONAL GAUSSIAN APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-adaptive detection of LLM-generated text aims to learn from human and machine passages a scoring rule for a specified deployment criterion. We formulate detection as a hypothesis-testing problem that maximizes machine-text detection power while controlling the false-positive rate on human writing. Existing logit-based detectors construct evidence from token log probability without accounting for contextual uncertainty, even though otherwise similar token probabilities can provide different evidence under different prefixes. We therefore introduce an entropy-conditioned detection method that jointly models token log probability and contextual entropy through a two-dimensional tensor-spline witness and aggregates the resulting evidence into a passage-level score. The central challenge is learning this witness for the fixed-FPR power objective, whose direct formulation leads to an ill-posed optimization problem. We address this challenge by analyzing how token-level evidence accumulates across a passage and using this sequential structure to approximate the human and machine scores as Gaussian distributions. This reformulation converts the original objective into a tractable problem with a regularized closed-form solution and shows that fixed-FPR power, fixed-FNR human TNR, and AUC share the same optimization direction. Across four source models and five datasets, our method achieves relative AUROC-error reductions of 7.4% and 10.9% over the previous state-of-the-art method in white-box and strict proxy-model black-box evaluations, respectively.

1 INTRODUCTION

Large language models (LLMs) produce fluent text at a scale that benefits programming, research, and creative work, but can also be used to produce misleading synthetic content (Zellers et al., 2019). In moderation or academic-integrity settings, detection is not merely a ranking problem: false positives can unfairly penalize human authors, with documented disparities for non-native English writers (Liang et al., 2023). We therefore seek high machine-text detection power subject to an explicit upper bound on the human false-positive rate (FPR).

Conditional-probability detectors use the probability structure of a pretrained language model without training a task-specific semantic classifier. DetectGPT (Mitchell et al., 2023) compares a passage with perturbations, and Fast-DetectGPT (Bao et al., 2024) replaces repeated perturbation with conditional token comparison. These methods use fixed score transformations. AdaDetectGPT (Zhou et al., 2025) makes the comparison data-adaptive by learning from external text a one-dimensional witness of token log probability. This progression raises a more general question: how should **a witness be represented and fitted** when its input is a dependent token sequence but its objective is passage-level class separation?

In answering the question, two issues motivate our reformulation. First, log probability alone does not fully describe the token's contribution to the conditional-probability detectors. Equal observed-token probabilities can occur when the remaining probability mass is either concentrated or diffuse. A one-dimensional witness could assign the same response to these cases and therefore cannot distinguish their different uncertainty profiles. Moreover, the passage-specific variance divisor used by conditional-discrepancy scores is computed under the accessible reference law, and thus the

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

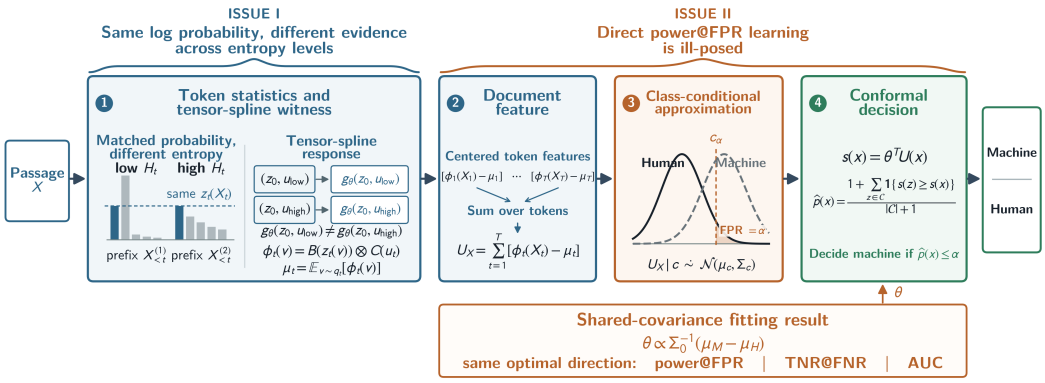

Figure 1: Overview of our method. An entropy-conditioned tensor-spline converts token log probabilities and prefix-entropy percentiles into centered features, which are summed as U_X. A class-conditional Gaussian approximation gives the fitted score direction; independent human passages then calibrate the frozen detector.

variance may not be at the actual human or machine fluctuation scale, leading to shifts in passage rankings. Second, fitting a data-adaptive witness function requires a valid link from dependent token increments to the class-conditional score distributions. The lower-bound derivation used for AdaDetectGPT (Zhou et al., 2025) reverses the relevant Jensen inequality; its closed-form direction therefore does not follow from the claimed TNR lower bound. Section 3 isolates this step and Appendix A.1 gives a counterexample.

We start from a hypothesis test that fixes the human false-positive rate (FPR). To construct its score, we use entropy to quantify contextual uncertainty and learn an entropy-conditioned witness of token log probability represented by a tensor-product spline. The sequential responses are accumulated in an unnormalized linear feature sum, whose class-conditional distribution we then model separately for human and machine text. We establish conditions under which this passage feature admits a class-conditional Gaussian approximation. We then prove that the true-positive rate (TPR) at fixed FPR, the human true-negative rate (TNR) at fixed machine false-negative rate (FNR), and the area under the ROC curve (AUC) are all monotone in the same standardized mean gap and hence select the same inverse-covariance projection direction. Regularized estimation fits this direction, and an independent human calibration set supplies the final finite-sample FPR guarantee. Our method pipeline is demonstrated in Figure 1.

Experiments on four source models and five datasets show that our method behaves better than existing logit-based detectors. Given fixed FPRs controlled by independent human calibration, our method achieves relative AUROC-error reductions of 7.4% and 10.9% over the state-of-the-art method in white-box and strict proxy-model black-box evaluations, respectively.

Our main theoretical contribution is a reformulation of the data-adaptive witness-fitting problem for a fixed-human-FPR test. We model contextual evidence with an entropy-conditioned tensor-spline witness of log probability. To fit the witness, we decompose the document feature sum into predictable drift and martingale fluctuation under each passage class, yielding explicit conditions for a class-conditional Gaussian approximation. We also prove that fixed-FPR power, fixed-FNR human TNR, and AUC have a common optimization direction under our assumptions, which can be efficiently estimated by a regularized closed-form solution.

2 Problem Formulation

Let a passage be a token sequence $X = (X_1, \ldots, X_T)$. Denote the distribution of human-written passages by P and that of passages generated by a target LLM by Q. A detector $\varphi : \mathcal{X} \to \{0, 1\}$ returns one for 'machine generation' and zero for 'human writing.' At a prescribed human false-

positive rate α, our primary objective is

$$\max_{\varphi} \underbrace{\mathbb{P}_{X\sim Q}\{\varphi(X)=1\}}_{\text{TPR}} \quad \text{s.t.} \quad \underbrace{\mathbb{P}_{X\sim P}\{\varphi(X)=1\}}_{\text{FPR}} \le \alpha. \tag{1}$$

This is the Neyman–Pearson fixed-size detection problem (Neyman & Pearson, 1933). At fixed human FPR, maximizing the machine true-positive rate (TPR) is maximizing the detection power. We use this criterion to state the problem; the fitting analysis later shows when it shares an optimal score direction with the dual fixed-machine-FNR criterion and with AUROC.

Here we define the kind of models that will be used in our problem. At token position t, let $q_t(v \mid X_{<t})$ be a sampling/reference next-token distribution and let $r_t(v \mid X_{<t})$ be a scoring distribution. The reference model defines which alternative tokens are compared at the observed prefix; the scoring model assigns the log probabilities used by the detector. In the white-box setting, the target generator is available and serves both roles. In the black-box setting, the target generator and its logits are unavailable, so accessible proxy models supply q_t and r_t. No query to the target generator is required at detection time in the black-box setting.

Learning the detector in (1) is nevertheless nontrivial. Once φ is implemented by thresholding a learned score, both the score distribution and the threshold required to attain FPR α depend on the witness parameters. The human and machine generation laws are unknown, the empirical indicator objective is discontinuous, and a flexible witness can overfit a finite train set. A useful fitting principle must therefore replace this direct, ill-posed optimization with a tractable population criterion while preserving its fixed-human-FPR meaning. The next section reviews the existing relaxation and identifies why its justification is insufficient; Section 4 then derives our alternative from class-conditional score approximations.

3 Background and Motivation

3.1 From Conditional Comparison to an Adaptive Witness

For a candidate token v at position t, let $z_t(v) = \log r_t(v \mid X_{<t})$ be its scoring-model log probability. Fast-DetectGPT compares the observed token with alternatives under the same prefix:

$$T_{\text{Fast}}(X) = \frac{\sum_{t=1}^{T}[z_t(X_t) - \mathbb{E}_{v\sim q_t} z_t(v)]}{\sqrt{\sum_{t=1}^{T} \operatorname{Var}_{v\sim q_t}[z_t(v)]}}. \tag{2}$$

Conditional centering removes the reference evidence expected at that prefix; the denominator expresses the discrepancy in reference-standard-deviation units (Bao et al., 2024). AdaDetectGPT replaces z by a learned smooth one-dimensional witness w:

$$T_w(X) = \frac{\sum_{t=1}^{T}[w(z_t(X_t)) - \mathbb{E}_{v\sim q_t} w(z_t(v))]}{\sqrt{\sum_{t=1}^{T} \operatorname{Var}_{v\sim q_t}[w(z_t(v))]}}. \tag{3}$$

Unlike the fixed identity transformation in Fast-DetectGPT, this witness is estimated from external passages. AdaDetectGPT is therefore a data-adaptive detector: the labeled training distribution determines how log probability is mapped to evidence. AdaDetectGPT chooses to optimize the witness function by maximizing human true-negative rate (TNR) for a fixed machine false-negative rate (FNR) β. A machine-side martingale normal approximation provides the corresponding threshold. Since the resulting TNR still depends on β, the method optimizes a lower bound of TNR—$h_\beta(r)$ instead, which is further turned into maximizing r.

Why the lower-bound argument must be rebuilt. The derivation correctly obtains the pointwise bound $\Phi(z_\beta + r) \ge h_\beta(r)$, but its next step replaces $\mathbb{E}h_\beta(R)$ by $h_\beta(\mathbb{E}R)$ where R is the corresponding random variable of r, in the direction required to obtain the claimed population lower bound. This step is invalid: because h_β is concave in r, Jensen's inequality gives $\mathbb{E}h_\beta(R) \le h_\beta(\mathbb{E}R)$, the reverse direction. Consequently, the published derivation does not establish its final objective as a TNR lower bound. Algebraically solving that objective does not repair the missing inequality. Appendix A.1 gives the exact statement and a two-point counterexample. We instead return to

the original fixed-human-FPR problem and derive the fitting direction from class-conditional score distributions.

The denominator in (3) also has a narrower interpretation than a general class standard deviation. It is accumulated under the accessible reference law q_t; it matches the relevant martingale variance when the evaluated sequence is generated by that same law and the central limit conditions hold. However, human text and black-box target text generally do not satisfy this premise. Besides, since the divisor varies across passages, for optimization it is not a global rescaling and may change their ranking. We therefore retain the unnormalized reference-centered sum and learn its class-conditional fluctuation scale from training passages. Appendix A.2 states the exact-reference boundary case.

3.2 Martingale Central-Limit Theory

The ordinary central-limit theorem for independent observations does not apply directly in text detection because every token depends on its prefix. Instead we consider a sequence of token contributions after removing their conditional means. Let $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$. For each candidate token v, let $\{a_t(v)\}_{t=1}^T$ be a time series whose value at position t is determined by the prefix, and define

$$D_t = a_t(X_t) - \mathbb{E}_{v\sim q_t} a_t(v) \tag{4}$$

when $X_t \mid \mathcal{F}_{t-1} \sim q_t$. Then $\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] = 0$, although the variables D_t remain sequentially dependent.

The martingale central-limit theorem requires two familiar controls. First, the sum of the conditional variances, $s_T^2 = \sum_t \mathbb{E}[D_t^2 \mid \mathcal{F}_{t-1}]$, must approach a stable scale. Second, no single token or short span may account for a non-vanishing part of that scale. The latter requirement is expressed by

$$\frac{1}{s_T^2} \sum_{t=1}^{T} \mathbb{E}\big[D_t^2 \mathbf{1}\{|D_t| > \varepsilon s_T\} \mid \mathcal{F}_{t-1}\big] \xrightarrow{p} 0 \quad (\varepsilon > 0). \tag{5}$$

When these two requirements hold, the normalized sum of the token contributions converges to a standard Gaussian distribution:

$$\frac{\sum_{t=1}^T D_t}{\sqrt{\sum_{t=1}^T \mathbb{E}[D_t^2 \mid \mathcal{F}_{t-1}]}} \Rightarrow \mathcal{N}(0,1). \tag{6}$$

We use the standard formulation of McLeish (1974) and Hall & Heyde (1980). In our detector, the human or machine next-token law need not equal the reference distribution q_t. Consequently, subtracting the reference expectation does not necessarily give a conditionally zero-mean quantity. Section 4.2 therefore separates the document feature into a class-dependent mean component and a zero-mean fluctuation component. The theorem above is applied only to the latter. Appendix A.3 gives the complete multivariate conditions and proof. The Gaussian approximation is used to learn the witness, whereas the final false-positive guarantee is supplied by human calibration.

4 Methodology

4.1 Entropy-Conditioned Tensor-Spline Score

At position t, the scoring-model log probability is $z_t(v) = \log r_t(v \mid X_{<t})$. We summarize uncertainty in the complete reference distribution by $H_t = -\sum_v q_t(v \mid X_{<t}) \log q_t(v \mid X_{<t})$, and map it through a train-set empirical CDF to the frozen percentile $u_t = \widehat{F}_H(H_t) \in [0,1]$. Thus $z_t(v)$ measures how strongly the scoring model favors candidate v, while u_t identifies how uncertain the prefix is relative to training contexts. Calibration and test passages do not affect this transform.

The same observed-token log probability can arise when the remaining probability mass is concentrated on a few alternatives or spread over many. A one-dimensional function of z cannot distinguish these contexts. We therefore model their interaction with quadratic B-spline bases (de Boor, 2001) $B(z) = (B_1(z), \ldots, B_J(z))^\top$ and $C(u) = (C_1(u), \ldots, C_K(u))^\top$:

$$g_\theta(z,u) = \sum_{j=1}^{J}\sum_{k=1}^{K} \theta_{jk} B_j(z) C_k(u) = \theta^\top \phi(z,u), \qquad \phi(z,u) = B(z) \otimes C(u) \in \mathbb{R}^{JK}. \tag{7}$$

The tensor product lets the response to log probability vary smoothly with prefix uncertainty; setting that response constant in u recovers a one-dimensional witness.

For position t, write $\phi_t(v) = \phi(z_t(v), u_t)$ and $\mu_t = \mathbb{E}_{v\sim q_t}\phi_t(v)$. The centered vector $e_t(X) = \phi_t(X_t) - \mu_t$ compares the observed token with reference alternatives under the same prefix. Summing these comparisons gives the document feature and its scalar detection score:

$$U_X = \sum_{t=1}^{T} e_t(X) = \sum_{t=1}^{T} [\phi_t(X_t) - \mu_t], \qquad T_\theta(X) = \theta^\top U_X. \tag{8}$$

Training passages determine θ; after all feature transforms and the score direction are frozen, an independent human calibration set determines the decision threshold. Appendix A.4 gives the feature-space derivation and sampled-moment form; Appendix A.5 gives the entropy-independent special cases.

4.2 Class-Conditional Gaussian Approximation

The vectors $e_t(X)$ in (8) are not independent because each token and its feature vector depend on the preceding prefix. We therefore analyze their sum in the order in which the tokens occur. Fix a class $c \in \{H, M\}$, where H denotes human text and M denotes machine text, and let $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$ contain the tokens observed through position t. For each position, the centered token vector is decomposed into its conditional mean and the remaining part:

$$b_{c,t} = \mathbb{E}_c[e_t(X) \mid \mathcal{F}_{t-1}], \quad \eta_{c,t} = e_t(X) - b_{c,t}. \tag{9}$$

By construction, $\mathbb{E}_c[\eta_{c,t} \mid \mathcal{F}_{t-1}] = 0$. Summing the two parts over the document gives

$$U_X = \underbrace{\sum_{t=1}^{T} b_{c,t}}_{B_{c,T}} + \underbrace{\sum_{t=1}^{T} \eta_{c,t}}_{M_{c,T}}. \tag{10}$$

The first term $B_{c,T}$ sums up the conditional mean evidence associated with class c; differences between the human and machine values of this term create the detection signal. $M_{c,T}$ is the sum of the zero-mean deviations representing the remaining token-sequence variation.

Here, the accumulated drift $B_{c,T}$ is still random because its summands depend on the prefixes. We therefore separate its deterministic expectation from its passage-to-passage fluctuation. Since $\mathbb{E}_c M_{c,T} = 0$, define

$$m_{c,T} := \mathbb{E}_c B_{c,T} = \mathbb{E}_c U_X, \qquad R_{c,T} := B_{c,T} - m_{c,T}, \tag{11}$$

so that $B_{c,T} = m_{c,T} + R_{c,T}$. Here $m_{c,T}$ is the average class-dependent centered-evidence, whereas $R_{c,T}$ describes how that accumulated score changes across passages.

Lemma 1 (Zero-mean representation of the drift fluctuation) *Assume $\mathcal{F}_0$ is trivial and define*

$$\xi_{c,t,T} := \mathbb{E}_c[B_{c,T} \mid \mathcal{F}_t] - \mathbb{E}_c[B_{c,T} \mid \mathcal{F}_{t-1}]. \tag{12}$$

Then

$$\mathbb{E}_c[\xi_{c,t,T} \mid \mathcal{F}_{t-1}] = 0, \qquad R_{c,T} = \sum_{t=1}^{T} \xi_{c,t,T}, \qquad \mathbb{E}_c R_{c,T} = 0. \tag{13}$$

Thus the fluctuation of $B_{c,T}$ is also a sum of zero-mean sequential increments. Intuitively, $\xi_{c,t,T}$ measures how observing token X_t changes the expected drift accumulated over the rest of the passage. The full proof, based on the tower property and a telescoping sum, is given in Appendix A.3.

We can now combine the two random terms in the document feature. Define

$$\zeta_{c,t,T} := \eta_{c,t} + \xi_{c,t,T}, \qquad Z_{c,T} := M_{c,T} + R_{c,T} = \sum_{t=1}^{T} \zeta_{c,t,T}. \tag{14}$$

Both components of $\zeta_{c,t,T}$ have conditional mean zero, so the combined increments form a martingale-difference array. Let $A_{c,T} := \mathrm{Var}_c(U_X) \succ 0$ denote their deterministic total covariance. A martingale central-limit theorem applies when their accumulated conditional covariance approaches $A_{c,T}$ and no token or short prefix contributes to a non-vanishing share of the total variation. Appendix A.3 states these conditions and proves the following result.

Lemma 2 (Gaussian limit of the combined fluctuation) *Under the combined quadratic-variation and Lindeberg conditions in Appendix A.3,*

$$A_{c,T}^{-1/2} Z_{c,T} \Rightarrow \mathcal{N}(0, I), \qquad \textit{or equivalently} \qquad Z_{c,T} \mathrel{\dot\sim} \mathcal{N}(0, A_{c,T}). \tag{15}$$

Finally, substituting $B_{c,T} = m_{c,T} + R_{c,T}$ into (10) gives the exact identity

$$U_X - m_{c,T} = R_{c,T} + M_{c,T} = Z_{c,T}. \tag{16}$$

Subtracting the class location therefore turns the document feature exactly into the combined fluctuation. Combining the zero-mean sequential representation in Lemma 1 with the Gaussian limit in Lemma 2 yields our first main result.

Theorem 1 (Class-conditional Gaussian approximation) *Under the conditions of Lemmas 1 and 2, for each class $c \in \{H, M\}$,*

$$A_{c,T}^{-1/2}(U_X - m_{c,T}) \Rightarrow \mathcal{N}(0, I), \qquad U_X \mid c \mathrel{\dot\sim} \mathcal{N}(m_{c,T}, A_{c,T}). \tag{17}$$

For the fixed length regime used to fit the detector, we suppress T and write these two parameters as μ_c and Σ_c.

4.3 Fixed-FPR Optimization and the Common Direction

Theorem 1 supplies separate human and machine Gaussian approximations. Theorems 2–3 further assume equal class covariances for closed-form fitting.

Theorem 2 (Fixed-FPR optimal direction) *Under Theorem 1, suppose the human and machine document features have equal covariance $\Sigma_H = \Sigma_M =: \Sigma_0 \succ 0$. Every maximizer of fixed-human-FPR power then has direction*

$$\theta^\star \propto \Sigma_0^{-1}(\mu_M - \mu_H), \qquad (\theta^\star)^\top(\mu_M - \mu_H) > 0. \tag{18}$$

Here, let $\delta = \mu_M - \mu_H$, orient θ so that $\theta^\top \delta > 0$, and classify a passage as machine generated when $T_\theta(X) > c$. According to Theorem 1, projection onto θ, together with the equal covariance condition, gives

$$T_\theta \mid H \mathrel{\dot\sim} \mathcal{N}(\theta^\top \mu_H, \sigma_\theta^2), \qquad T_\theta \mid M \mathrel{\dot\sim} \mathcal{N}(\theta^\top \mu_M, \sigma_\theta^2), \quad \sigma_\theta^2 = \theta^\top \Sigma_0 \theta. \tag{19}$$

For a prescribed human FPR α, the threshold must satisfy

$$\mathbb{P}_H\{T_\theta > c\} = 1 - \Phi\left(\frac{c - \theta^\top \mu_H}{\sigma_\theta}\right) \le \alpha \quad \Longleftrightarrow \quad c \ge \theta^\top \mu_H + z_{1-\alpha}\sigma_\theta. \tag{20}$$

where $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. Raising c reduces the machine TPR, so the most powerful feasible test uses the smallest threshold allowed by (20):

$$c_\alpha(\theta) = \theta^\top \mu_H + z_{1-\alpha}\sigma_\theta. \tag{21}$$

Substituting this threshold into the machine tail probability gives

$$\mathrm{TPR}_\alpha(\theta) = \mathbb{P}_M\{T_\theta > c_\alpha(\theta)\} = \Phi\left(\frac{\theta^\top(\mu_M - \mu_H)}{\sqrt{\theta^\top \Sigma_0 \theta}} - z_{1-\alpha}\right). \tag{22}$$

Define the standardized class separation

$$R(\theta) = \frac{\theta^\top \delta}{\sqrt{\theta^\top \Sigma_0 \theta}}. \tag{23}$$

Equation (22) becomes $\mathrm{TPR}_\alpha(\theta) = \Phi(R(\theta) - z_{1-\alpha})$. Since Φ is increasing and $z_{1-\alpha}$ is fixed, maximizing power at FPR α is exactly the same as maximizing $R(\theta)$.

Appendix A.6 proves that this direction maximizes $R(\theta)$; it is the classical shared-covariance discriminant direction (Fisher, 1936).

The next result connects this direction to the other evaluation criteria.

Theorem 3 (Common direction for operating criteria) *Under Theorem 1, suppose* $\Sigma_H = \Sigma_M =: \Sigma_0 \succ 0$. *Fixed-human-FPR power, human TNR at fixed machine FNR, and AUC are then increasing functions of* $R(\theta)$ *and share the optimal direction* $\theta^\star \propto \Sigma_0^{-1}(\mu_M - \mu_H)$.

According to Theorem 1, the two class projections are Gaussian; Theorem 3's covariance condition makes the same $R(\theta)$ govern both remaining criteria. If machine FNR is fixed at β, write $z_\beta = \Phi^{-1}(\beta)$. The corresponding machine-score quantile and its human TNR are

$$d_\beta(\theta) = \theta^\top \mu_M + z_\beta \sigma_\theta, \qquad \mathrm{TNR}_\beta(\theta) = \mathbb{P}_H\{T_\theta \le d_\beta(\theta)\} = \Phi(R(\theta) + z_\beta). \tag{24}$$

For independent human and machine documents, the score difference is Gaussian with mean $\theta^\top \delta$ and variance $2\theta^\top \Sigma_0 \theta$, so

$$\mathrm{AUC}(\theta) = \mathbb{P}\{T_\theta(X_M) > T_\theta(X_H)\} = \Phi(R(\theta)/\sqrt{2}). \tag{25}$$

The three criteria remain different quantities: in particular, once human FPR is fixed at α, human TNR is already $1 - \alpha$ and is not another objective to optimize.

For finite-sample fitting, train passages provide the class means $\overline{U}_H, \overline{U}_M$, their difference $\widehat{\delta} = \overline{U}_M - \overline{U}_H$, and a regularized pooled covariance matrix $A \succ 0$. Covariance shrinkage, tensor-surface smoothness, and ridge regularization stabilize the discriminant fit (Friedman, 1989; Hastie et al., 1995) and give the direction $\widetilde{\theta} = A^{-1}\widehat{\delta}$. Here surface-smoothness selection uses training-only cross-validation to choose the strength of a second-difference penalty along both axes of the tensor-spline coefficient surface. Covariance-shrinkage selection preserves featurewise variances while damping noisy empirical correlations. All hyperparameters, features and the scores are selected on training data and frozen before calibration. Appendix A.7 gives the full estimator, and Algorithm 1 gives the complete procedure.

Calibration. Let $C_1, \ldots, C_m$ be independent human calibration passages. After the score is frozen, the query x receives the conformal rank value (Vovk et al., 2005)

$$\widehat{p}(x) = \frac{1 + \sum_{i=1}^m \mathbf{1}\{T_{\widehat{\theta}}(C_i) \ge T_{\widehat{\theta}}(x)\}}{m+1}.$$

The final decision rule is $\widehat{\varphi}_\alpha(x) = \mathbf{1}\{\widehat{p}(x) \le \alpha\}$: a value of one declares x machine generated, while zero declares it human. Thus a passage is flagged only when few calibration human passages receive scores at least as large as its score. Appendix C states the required calibration-to-test condition and proves finite-sample human-FPR control.

5 Experiments

We evaluate white-box and proxy-model black-box detection through overall ranking, low-FPR performance, and Gaussian-fit diagnostics. Configuration-level results and implementation details appear in the appendix.

Data, models, and splits. We evaluate 20 configurations spanning five datasets—C4 (Raffel et al., 2020), PubMedQA (PubMed in tables) (Jin et al., 2019), SQuAD (Rajpurkar et al., 2016), WritingPrompts (Fan et al., 2018), and XSum (Narayan et al., 2018)—and four source models: LLaMA-2-7B (Touvron et al., 2023), OPT-2.7B/6.7B (Zhang et al., 2022), and Falcon-7B (Penedo et al., 2023). Each configuration uses 500 training pairs, 500 independent human calibration passages, and disjoint test sets of 500 human and 500 machine passages. We prompt with 120 tokens and generate 150–200 tokens using top-$k = 50$ sampling at temperature 1. Passages are length matched; source overlap, normalized duplicates, and prefix collisions are removed. Test data are never used for fitting, selection, orientation, or calibration; Appendix B gives further details.

Baselines and evaluation. White-box baselines are likelihood, entropy, rank, and log-rank (Gehrmann et al., 2019; Mitchell et al., 2023), LRR (Su et al., 2023), Fast-DetectGPT (Bao et al., 2024), and AdaDetectGPT (Zhou et al., 2025). In black-box evaluation, no method accesses source logits; GPT-J-6B (Wang & Komatsuzaki, 2021) is the reference model and GPT-Neo-2.7B (Black

Table 1: White-box independent full-text AUROC. Dataset columns average four source models; Macro averages all 20 configurations. Ours uses $T_{\widehat{\theta}}$.

Method	C4	PubMed	SQuAD	Writing	XSum	Macro
Likelihood	0.5707	0.5727	0.5770	0.6714	0.5537	0.5891
Entropy	0.5500	0.5575	0.5474	0.5303	0.5761	0.5523
Rank	0.5929	0.5919	0.5872	0.6193	0.5777	0.5938
LogRank	0.5951	0.6014	0.5984	0.6909	0.5779	0.6127
LRR	0.6553	0.6635	0.6473	0.7134	0.6215	0.6602
Fast-DetectGPT	0.7681	0.7502	0.7591	0.8589	0.7289	0.7730
AdaDetectGPT	0.8253	0.7987	0.8157	0.9026	0.7797	0.8244
Ours	**0.8447**	**0.8087**	**0.8261**	**0.9120**	**0.7956**	**0.8374**

Table 2: Black-box independent full-text AUROC with fixed proxy models. Dataset columns average four source models; Macro averages all 20 configurations.

Method	C4	PubMed	SQuAD	Writing	XSum	Macro
Fast-DetectGPT	0.7034	0.5763	0.6939	0.7993	0.6477	0.6841
AdaDetectGPT	0.7623	0.6492	0.7444	**0.8419**	0.7129	0.7421
Ours	**0.7776**	**0.7176**	**0.7773**	0.8404	**0.7377**	**0.7701**

et al., 2021) the scorer. AdaDetectGPT and our method train on the other four datasets, tune by leave-one-training-dataset-out validation, and refit before held-out-dataset testing. We report AUROC, standardized pAUC over 0–5% FPR (McClish, 1989), and power at 1% and 5% human FPR. All methods use identical full passages and independent human calibration sets; Appendix B details baseline configurations.

White-box detection. Table 1 shows that our method achieves the highest AUROC on every dataset and in all 20 model–dataset configurations. Its macro AUROC is 0.8374, compared with 0.8244 for AdaDetectGPT and 0.7730 for Fast-DetectGPT. The gain over AdaDetectGPT amounts to a 7.4% relative reduction in AUROC error and ranges from 0.0094 on WritingPrompts to 0.0194 on C4 across datasets. A paired-cell bootstrap (Efron, 1979) gives a 95% interval of $[0.0101, 0.0160]$, and a one-sided Wilcoxon signed-rank test (Wilcoxon, 1945) gives $p = 9.54 \times 10^{-7}$. Per-model results appear in Table 3.

Black-box detection. Table 2 reports results with fixed proxy models. Our method reaches a macro AUROC of 0.7701, exceeding AdaDetectGPT by 0.0280 and Fast-DetectGPT by 0.0860. The former gain is a 10.9% relative reduction in AUROC error, with wins in 17 of 20 configurations. Improvements are largest on PubMed and SQuAD; WritingPrompts is the only dataset-average exception, where AdaDetectGPT is higher by 0.0015. This near tie contrasts with our 0.0094 white-box gain on the same dataset and may reflect proxy-model mismatch: open-ended WritingPrompts continuations have greater stylistic and prefix-level variability, making the learned score less transferable. The paired bootstrap interval against AdaDetectGPT is $[0.0169, 0.0411]$, with one-sided Wilcoxon $p = 2.38 \times 10^{-5}$. Table 4 gives all proxy baselines and individual configurations; Appendix D.3 gives matched component factorials under the primary score.

Performance at fixed false-positive rates. Figure 2 compares the three conditional detectors over the complete 0–5% FPR range. Each curve is obtained by iterating over test-score thresholds within every configuration, computing the resulting empirical FPR and TPR, interpolating to a common FPR grid, and then averaging the same 20 model–dataset configurations. Calibration scores are used only for the hollow-diamond operating points.

Our curve remains above both baselines throughout this range. At empirical FPRs of 1% and 5%, its macro powers are 21.22% and 43.06%, compared with 17.47%/38.35% for AdaDetectGPT and 10.77%/27.52% for Fast-DetectGPT. After independent calibration, our method attains 17.72%/40.02% power at realized FPRs of 0.84%/4.40%; AdaDetectGPT attains 12.93%/35.92%, and Fast-DetectGPT 8.81%/27.12%. The corresponding standardized pAUC@5% values are

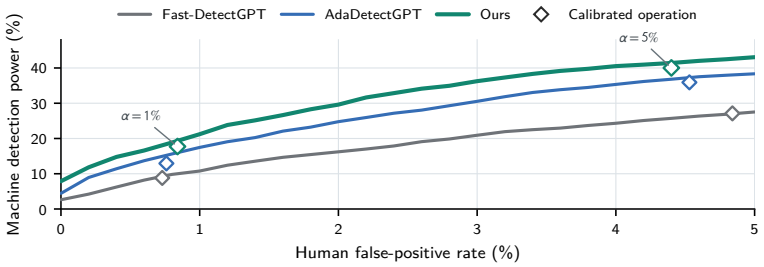

Figure 2: White-box ROC comparison over 0–5% FPR. Curves show mean empirical TPR across 20 model–dataset configurations. Hollow diamonds mark mean test-set FPR–TPR pairs after independent human calibration at nominal $\alpha = 1\%$ and 5%.

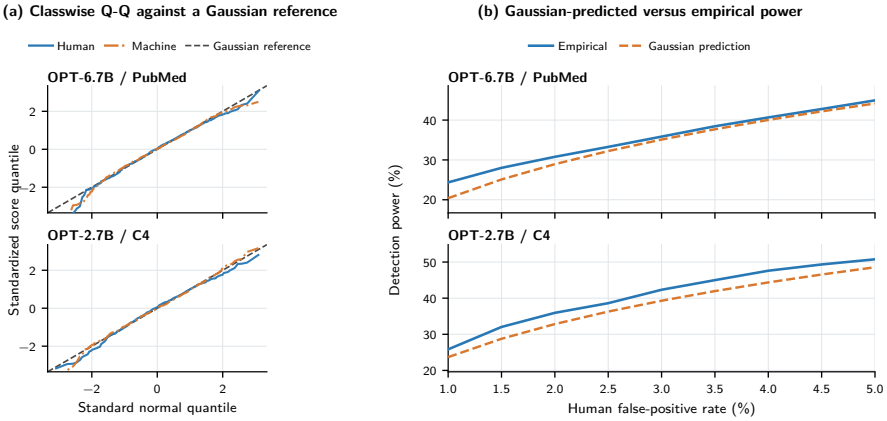

Figure 3: Finite-sample evaluation of the Gaussian working model. Each row is one model–dataset configuration. Left: standardized human and machine score quantiles against a Gaussian reference. Right: Gaussian-predicted versus held-out empirical power. Complete normality tests appear in Appendix D.2.

0.6424, 0.6183, and 0.5751. Black-box operating points and complete white-box results appear in Appendix D.1.

Gaussian working-model diagnostics. Figure 3 tests the approximation used to fit the projection direction through classwise Q–Q plots and held-out power prediction.

In the displayed configurations, both Q–Q curves remain close to their Gaussian references, and predicted power follows held-out power over the 1–5% FPR range. Across all 20 configurations, mean absolute power error is 0.0393 at 1% FPR and 0.0288 at 5% FPR, while predicted and empirical AUROC differ by 0.0043 on average. Fitted-normal tests reject none of the 20 machine-score distributions at the 5% level, although several human distributions deviate detectably from exact normality. Yet the approximation captures their separation well enough to support direction fitting and useful power prediction.

6 Conclusion

We introduce an entropy-conditioned tensor-spline detector that combines token log probability and contextual uncertainty at fixed human FPR. A martingale decomposition yields a Gaussian working model and, under shared covariance, one closed-form direction for fixed-FPR power, fixed-FNR human TNR, and AUC. Conformal calibration controls finite-sample FPR without exact normality. Across four models and five datasets, our method reduces AUROC error over AdaDetectGPT by 7.4% (white-box) and 10.9% (strict proxy black-box).

AI USE STATEMENT

We used OpenAI's ChatGPT and Codex during the development and preparation of this work. For research ideation and execution, these tools were used to discuss and refine the entropy-conditioned detection formulation, propose diagnostic and ablation analyses, inspect existing experimental outputs, and assist with the implementation and debugging of analysis and plotting code. In particular, AI assistance was used in preparing code for score reconstruction, fixed-FPR evaluation, token-dominance diagnostics, and evaluation of the class-conditional Gaussian working model.

For mathematical development, the tools were used to examine intermediate derivations, help check and polish the proof of the drift–fluctuation decomposition, and review the parts of the arguments concerning the common optimal projection direction for fixed-FPR power, fixed-FNR human TNR, and AUC. The main ideas and key facts were initiated and proved by the authors. All assumptions, derivations, and final mathematical claims were independently checked by the authors. AI-generated suggestions were not treated as valid proofs without independent verification.

The tools were also used to identify potentially relevant literature, generate preliminary drafts of selected passages, improve readability, revise figure captions and algorithm descriptions, and assist with LaTeX formatting. Candidate references suggested by AI were checked against the original papers. AI-assisted code was executed and checked against stored scores and experimental outputs. All numerical results reported in the paper were obtained from the authors' executed experiments rather than generated or estimated by an AI system. The authors reviewed all AI-assisted text, code, mathematical arguments, citations, figures, and numerical summaries, and take responsibility for the final content of the paper.

ETHICS STATEMENT

Machine-generated text detectors can produce false accusations, particularly under distribution shift or after adversarial rewriting. We therefore frame the task around explicit false-positive control, report realized false-positive rates together with detection power, and discuss the limits of exchangeability and cross-dataset generalization. The work uses existing text corpora and pretrained language models and does not involve new human-subject data collection.

REPRODUCIBILITY STATEMENT

The data splits, generation protocol, benchmark configurations, training and calibration order, and operating-point definitions are specified in Section 5 and Appendix B. The assumptions, derivations, and proofs supporting the method are collected in the remaining appendix sections. These descriptions separate train-only selection, independent human calibration, and final test evaluation so that the reported protocol can be reproduced without using test labels for fitting or orientation.

REFERENCES

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations*, 2024.

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*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with Mesh-TensorFlow. EleutherAI, 2021.

Carl de Boor. *A Practical Guide to Splines*. Springer, New York, revised edition, 2001.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121, 1996. doi: 10.1214/ss/1038425655.

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

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936.

Jerome H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989. doi: 10.1080/01621459.1989.10478752.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, 2019.

Peter Hall and Christopher C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, New York, 1980.

Trevor Hastie, Andreas Buja, and Robert Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995.

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

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. GPT detectors are biased against non-native english writers. *Patterns*, 4(7):100779, 2023. doi: 10.1016/j.patter.2023.100779.

Hubert W. Lilliefors. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967. doi: 10.1080/01621459.1967.10482916.

Donna Katzman McClish. Analyzing a portion of the ROC curve. *Medical Decision Making*, 9(3): 190–195, 1989. doi: 10.1177/0272989X8900900307.

Donald L. McLeish. Dependent central limit theorems and invariance principles. *The Annals of Probability*, 2(4):620–628, 1974.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 24950–24962, 2023.

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

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231:289–337, 1933.

Guilherme Penedo et al. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2023.

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

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

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

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 billion parameter autoregressive language model. EleutherAI, 2021.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

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*, volume 32, 2019.

Susan Zhang et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

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

A Theoretical Details and Proofs

A.1 Failure of AdaDetectGPT Lemma S2

We analyze the exact statement of Lemma S2 in Zhou et al. (2025). For every scalar random variable R and $0 < \beta < 1/2$, it claims

$$\mathbb{E}\Phi(z_\beta + R) \geq \min\{1 - \beta,\ \beta + \phi(z_\beta)\mathbb{E}R\}. \tag{26}$$

Define $h_\beta(r) = \min\{\beta + \phi(z_\beta)r, 1 - \beta\}$. Its affine and constant branches meet at $r_\star = (1 - 2\beta)/\phi(z_\beta) > 0$. Hence $h_\beta(r) = \beta + \phi(z_\beta)r$ for $r \leq r_\star$ and $h_\beta(r) = 1 - \beta$ for $r \geq r_\star$. Its left and right derivatives at the only kink are $\phi(z_\beta) > 0$ and 0, respectively. A continuous piecewise-affine function whose slope decreases at every kink is concave. Jensen's inequality therefore gives

$$\mathbb{E}h_\beta(R) \leq h_\beta(\mathbb{E}R), \tag{27}$$

the reverse of the final inequality used to prove (26).

The pointwise argument preceding that step remains valid: $\Phi(z_\beta + r) \geq h_\beta(r)$. Indeed, for $r \leq 0$, convexity of Φ on the negative half-line gives $\Phi(z_\beta + r) - \beta \geq \phi(z_\beta)r$. For $0 < r < z_{1-\beta} - z_\beta$, every intermediate point lies in $[z_\beta, z_{1-\beta}]$, where $\phi(\xi) \geq \phi(z_\beta)$. For larger r, $\Phi(z_\beta + r) \geq 1 - \beta$. Thus the valid chain stops at $\mathbb{E}\Phi(z_\beta + R) \geq \mathbb{E}h_\beta(R)$; concavity does not permit replacing the right side by $h_\beta(\mathbb{E}R)$.

To show directly that the stated lemma is false for every $0 < \beta < 1/2$, let $c_\beta = (1-2\beta)/\phi(z_\beta) > 0$ and choose $0 < p < (1 - 2\beta)/(1 - \beta)$. Set $R = 0$ with probability $1 - p$ and $R = c_\beta/p$ with probability p. Then $\mathbb{E}R = c_\beta$, so the right side of (26) is $h_\beta(\mathbb{E}R) = 1 - \beta$, whereas

$$\mathbb{E}\Phi(z_\beta + R) = (1 - p)\beta + p\Phi(z_\beta + c_\beta/p) < (1 - p)\beta + p = \beta + p(1 - \beta) < 1 - \beta. \tag{28}$$

This random variable satisfies the lemma's unrestricted quantifier but violates its conclusion. Therefore the final Jensen step, and not the earlier pointwise bound or the later linear-algebra solve, is the precise failure.

A.2 Why Passage-Specific Standardization Is Not Required

For a frozen direction, the expected linear score in class c is $\mathbb{E}_c T_\theta = \theta^\top \mathbb{E}_c B_{c,T}$; hence the difference of the class-conditional predictable drifts is already a location signal. The shared-covariance model measures this signal relative to its fluctuation scale through $R(\theta)$ while fitting the direction. Once θ is fixed, any positive global rescaling leaves ROC rankings and conformal ranks unchanged, and human calibration estimates the required threshold directly. There is therefore no probabilistic need to divide each passage by a different reference-law variance.

The usual standard-normal argument for S_θ has a much narrower scope. If an evaluated machine passage is generated exactly from the same q_t, then $b_{c,t} = 0$, so $B_{c,T} = m_{c,T} = R_{c,T} = 0$ and $\xi_{c,t,T} = 0$. Consequently,

$$\widetilde{Q}_{q,T} = \sum_t \mathbb{E}_q[(\phi_t(X_t) - \mu_t)(\phi_t(X_t) - \mu_t)^\top \mid \mathcal{F}_{t-1}] = V_X. \tag{29}$$

With a frozen θ, inactive numerical floor, and (38)–(39), this yields

$$S_\theta(X) = \frac{\theta^\top U_X}{\sqrt{\theta^\top V_X \theta}} \Rightarrow \mathcal{N}(0, 1). \tag{30}$$

Human text does not satisfy the exact-reference premise, and black-box machine text generally does not either. Then both the drift and true quadratic variation differ from the reference denominator. Because $X \mapsto \sqrt{\theta^\top V_X \theta}$ is passage dependent, S_θ is not a global monotone transform of T_θ and can change rankings. This is why we treat it as an empirical ablation rather than the default deployment score.

A.3 Random Drift Fluctuation and MCLT Conditions

Fix class $c \in \{H, M\}$ and let $p_{c,t}(\cdot \mid X_{<t})$ denote its true conditional token law. Since q_t is only the accessible reference law,

$$\mathbb{E}_c[\phi_t(X_t) - \mu_t \mid \mathcal{F}_{t-1}] = \mathbb{E}_{p_{c,t}}\phi_t(v) - \mathbb{E}_{q_t}\phi_t(v) = b_{c,t}.$$

Subtracting $b_{c,t}$ produces the martingale difference $\eta_{c,t} = \phi_t(X_t) - \mu_t - b_{c,t}$ used in (10). Consequently, $M_{c,T} = \sum_t \eta_{c,t}$ has mean zero and $m_{c,T} = \mathbb{E}_c U_X = \mathbb{E}_c B_{c,T}$, as used in (11). The remaining term $R_{c,T} = B_{c,T} - m_{c,T}$ need not be negligible. We first prove that it also has an exact zero-mean sequential representation.

Proof of Lemma 1. Define the Doob martingale

$$G_{c,t,T} := \mathbb{E}_c[B_{c,T} \mid \mathcal{F}_t], \qquad t = 0, \dots, T,$$

and let $\xi_{c,t,T} = G_{c,t,T} - G_{c,t-1,T}$, as in (12). By the tower property,

$$\mathbb{E}_c[\xi_{c,t,T} \mid \mathcal{F}_{t-1}] = \mathbb{E}_c[G_{c,t,T} \mid \mathcal{F}_{t-1}] - G_{c,t-1,T} = G_{c,t-1,T} - G_{c,t-1,T} = 0. \tag{31}$$

Because $B_{c,T}$ is $\mathcal{F}_T$-measurable and $\mathcal{F}_0$ is trivial, the increments telescope:

$$\sum_{t=1}^{T} \xi_{c,t,T} = G_{c,T,T} - G_{c,0,T} = B_{c,T} - \mathbb{E}_c B_{c,T} = R_{c,T}. \tag{32}$$

Taking expectations yields $\mathbb{E}_c R_{c,T} = 0$. If a fixed initial prompt is included in $\mathcal{F}_0$, the same proof holds conditionally, with $\mathbb{E}_c[B_{c,T} \mid \mathcal{F}_0]$ replacing $m_{c,T}$.

This representation also explains the meaning of each increment. Since $b_{c,s}$ is $\mathcal{F}_{s-1}$-measurable, the terms with $s \le t$ are already known before X_t is observed. Hence

$$\xi_{c,t,T} = \sum_{s=t+1}^{T} \left\{ \mathbb{E}_c[b_{c,s} \mid \mathcal{F}_t] - \mathbb{E}_c[b_{c,s} \mid \mathcal{F}_{t-1}] \right\}. \tag{33}$$

Thus $\xi_{c,t,T}$ records how observing token X_t changes the expected drift at future positions.

Combining the two fluctuations. Let $\zeta_{c,t,T} = \eta_{c,t} + \xi_{c,t,T}$ and $Z_{c,T} = \sum_t \zeta_{c,t,T}$ as in (14). Equation (31) and $\mathbb{E}_c[\eta_{c,t} \mid \mathcal{F}_{t-1}] = 0$ imply

$$\mathbb{E}_c[\zeta_{c,t,T} \mid \mathcal{F}_{t-1}] = 0. \tag{34}$$

Moreover, combining (10), (11), and (32) gives the exact identity

$$U_X - m_{c,T} = M_{c,T} + R_{c,T} = \sum_{t=1}^{T} \zeta_{c,t,T} = Z_{c,T}. \tag{35}$$

Therefore the complete class-centered feature sum, including random drift variation, is a martingale sum.

Its accumulated conditional covariance is

$$\widetilde{Q}_{c,T} := \sum_{t=1}^{T} \mathbb{E}_c[\zeta_{c,t,T}\zeta_{c,t,T}^{\top} \mid \mathcal{F}_{t-1}]. \tag{36}$$

Martingale differences at distinct times are orthogonal. Therefore, with $A_{c,T} := \mathrm{Var}_c(U_X)$,

$$\mathbb{E}_c \widetilde{Q}_{c,T} = \mathrm{Var}_c\left(\sum_{t=1}^{T} \zeta_{c,t,T} \right) = \mathrm{Var}_c(U_X) = A_{c,T}. \tag{37}$$

Unlike the conditional covariance of $\eta_{c,t}$ alone, $\widetilde{Q}_{c,T}$ includes local token variation, random drift influence, and their within-time interaction.

For the multivariate martingale CLT, assume $A_{c,T} \succ 0$ and require

$$A_{c,T}^{-1/2} \widetilde{Q}_{c,T} A_{c,T}^{-1/2} \xrightarrow{p} I. \tag{38}$$

Also, for every fixed $a \in \mathbb{R}^{JK}$ and every $\epsilon > 0$, require

$$\sum_{t=1}^{T} \mathbb{E}_c\left[\left(a^{\top} A_{c,T}^{-1/2} \zeta_{c,t,T}\right)^2 \mathbf{1}\left\{ \left| a^{\top} A_{c,T}^{-1/2} \zeta_{c,t,T} \right| > \epsilon \right\} \middle| \mathcal{F}_{t-1} \right] \xrightarrow{p} 0. \tag{39}$$

Proof of Lemma 2. Fix $a \neq 0$ and define

$$Y_{c,t,T} := \frac{a^\top A_{c,T}^{-1/2} \zeta_{c,t,T}}{\|a\|_2}.$$

By (34), these variables form a scalar martingale-difference array. Condition (38) gives

$$\sum_{t=1}^{T} \mathbb{E}_c[Y_{c,t,T}^2 \mid \mathcal{F}_{t-1}] = \frac{a^\top A_{c,T}^{-1/2} \widetilde{Q}_{c,T} A_{c,T}^{-1/2} a}{\|a\|_2^2} \xrightarrow{p} 1,$$

and (39), after rescaling ϵ, is its conditional Lindeberg condition. The scalar martingale CLT therefore yields

$$\frac{a^\top A_{c,T}^{-1/2} Z_{c,T}}{\|a\|_2} \Rightarrow \mathcal{N}(0,1). \tag{40}$$

Because this holds for every fixed direction a, the Cramér–Wold theorem gives $A_{c,T}^{-1/2} Z_{c,T} \Rightarrow \mathcal{N}(0, I)$, which proves Lemma 2. Finally, (35), together with Lemma 1, gives (17) and proves Theorem 1.

Condition (38) requires the total conditional covariance to stabilize, while (39) rules out a single token or short prefix contributing a nonvanishing share of the complete document fluctuation. Bounded local-support spline features help control the direct innovation $\eta_{c,t}$, but they do not by themselves bound one token's influence on all future drift terms through $\xi_{c,t,T}$. The combined conditions therefore additionally require sufficiently weak and stable long-range prefix influence. They support the Gaussian fitting approximation; they do not assert exact normality for every finite passage.

A.4 Feature-Space Form of the Tensor-Spline Score

First, the tensor notation is dimensionally explicit. Reshape the paired-index vector θ as the matrix Θ with $\Theta_{jk} = \theta_{(j,k)}$. Then

$$\theta^\top \phi(z,u) = \sum_{j=1}^{J} \sum_{k=1}^{K} \Theta_{jk} B_j(z) C_k(u) = B(z)^\top \Theta C(u),$$

which proves the second equality in (7). By this identity and linearity of conditional expectation,

$$\Delta_{\theta,t}(X) = \theta^\top \phi_t(X_t) - \theta^\top \mathbb{E}_{q_t} \phi_t(v) = \theta^\top [\phi_t(X_t) - \mu_t]. \tag{41}$$

Similarly,

$$\begin{aligned} \mathrm{Var}_{q_t}[\theta^\top \phi_t(v)] &= \mathbb{E}_{q_t}\left[\theta^\top (\phi_t - \mu_t)(\phi_t - \mu_t)^\top \theta\right] \\ &= \theta^\top \Sigma_t \theta. \end{aligned} \tag{42}$$

Summing (41) and (42) over t and using (8) yields

$$\sum_t \Delta_{\theta,t} = \theta^\top U_X, \qquad \sum_t s_{\theta,t}^2 = \theta^\top V_X \theta,$$

which establishes the primary score (8) and the variance term in the secondary ablation (55). The variance sum is nonnegative because every $\Sigma_t \succeq 0$, and the floor ε affects only numerically degenerate values of that secondary score.

For the sampled estimator, let $W_{t1}, \ldots, W_{tM} \overset{\text{iid}}{\sim} q_t$ and write $Y_{tm} = \phi_t(W_{tm})$. Define

$$\widehat{\mu}_t = \frac{1}{M} \sum_{m=1}^{M} Y_{tm}, \tag{43}$$

$$\widehat{\Sigma}_t = \frac{1}{M-1} \sum_{m=1}^{M} (Y_{tm} - \widehat{\mu}_t)(Y_{tm} - \widehat{\mu}_t)^\top. \tag{44}$$

Conditioned on $X_{<t}$, the standard sample-moment identities give $\mathbb{E}[\widehat{\mu}_t \mid X_{<t}] = \mu_t$. For the covariance, expand

$$\begin{aligned}\sum_{m=1}^{M}(Y_{tm}-\widehat{\mu}_t)(Y_{tm}-\widehat{\mu}_t)^\top &= \sum_{m=1}^{M}(Y_{tm}-\mu_t)(Y_{tm}-\mu_t)^\top \\ &\quad - M(\widehat{\mu}_t-\mu_t)(\widehat{\mu}_t-\mu_t)^\top.\end{aligned}$$

The two expectations on the right are $M\Sigma_t$ and Σ_t, respectively, so division by $M-1$ gives $\mathbb{E}[\widehat{\Sigma}_t \mid X_{<t}] = \Sigma_t$. Therefore

$$\widehat{U}_X = \sum_t [\phi_t(X_t) - \widehat{\mu}_t], \qquad \widehat{V}_X = \sum_t \widehat{\Sigma}_t$$

satisfy $\mathbb{E}[\widehat{U}_X \mid X] = U_X$ and $\mathbb{E}[\widehat{V}_X \mid X] = V_X$. This establishes conditional unbiasedness of the feature moments. The primary score $\widehat{T}_\theta = \theta^\top \widehat{U}_X$ is consequently conditionally unbiased for fixed θ. Because a ratio of random estimators is not generally an unbiased estimator of the corresponding ratio, no unbiasedness claim is made for the sampled self-normalized ablation.

The exact-reference boundary case is stated in Appendix A.2.

A.5 Entropy-Independent Witness Special Cases

For any coefficient vector β, the partition-of-unity property gives

$$B(z)^\top(\beta \mathbf{1}_K^\top)C(u) = B(z)^\top \beta\, \mathbf{1}_K^\top C(u) = B(z)^\top \beta.$$

Thus the tensor space contains every entropy-independent witness in the one-dimensional spline span. If $B(z)^\top \beta_{\mathrm{id}} = z$ on $\mathcal{I}$, substituting this witness into the primary score definition gives

$$T_{\mathrm{id}}(X) = \sum_t [z_t(X_t) - \mathbb{E}_{v\sim q_t} z_t(v)].$$

Using the corresponding conditional variance in the optional self-normalized score gives

$$S_{\mathrm{id}}(X) = \frac{\sum_t [z_t(X_t) - \mathbb{E}_{v\sim q_t} z_t(v)]}{\sqrt{\max\{\sum_t \mathrm{Var}_{v\sim q_t}[z_t(v)], \varepsilon\}}}, \tag{45}$$

the Fast-DetectGPT conditional probability discrepancy computed with the same reference q and scoring model r.

A.6 Proofs of Theorems 2 and 3

Section 4.3 derives the human-FPR threshold and the expressions for fixed-FPR power, fixed-FNR TNR, and AUC. It remains only to maximize the common quantity $R(\theta)$. Under the equal-covariance conditions stated directly in Theorems 2 and 3, $\Sigma_0 \succ 0$, so the following whitening transformation is well defined. Writing $a = \Sigma_0^{1/2}\theta$ and $b = \Sigma_0^{-1/2}\delta$, Cauchy–Schwarz gives

$$\frac{\theta^\top \delta}{\sqrt{\theta^\top \Sigma_0 \theta}} = \frac{a^\top b}{\|a\|_2} \le \|b\|_2,$$

with equality if and only if a is a positive multiple of b, or equivalently $\theta \propto \Sigma_0^{-1}\delta$. The sign condition $\theta^\top \delta > 0$ selects the upper-tail orientation; multiplying by a positive scalar leaves $R(\theta)$, all three criteria, and the induced ranking unchanged. This proves Theorems 2 and 3.

A.7 Train-Set Estimation and Regularized Optimization

Let $\mathcal{D}_{\mathrm{tr}}^c$ contain n_c train passages from class $c \in \{H, M\}$. The plug-in estimators used in every experiment are

$$\begin{aligned}\overline{U}_c &= \frac{1}{n_c}\sum_{X\in\mathcal{D}_{\mathrm{tr}}^c} U_X, & \widehat{\Sigma}_c &= \frac{1}{n_c-1}\sum_{X\in\mathcal{D}_{\mathrm{tr}}^c}(U_X-\overline{U}_c)(U_X-\overline{U}_c)^\top, \\ \widehat{\delta} &= \overline{U}_M - \overline{U}_H, & \widehat{\Sigma}_\Delta &= \widehat{\Sigma}_M + \widehat{\Sigma}_H.\end{aligned} \tag{46}$$

These train-set class means and covariance sums estimate the population mean gap and within-class covariance up to an irrelevant common scale. Direct inversion of $\widehat{\Sigma}_\Delta$ can be unstable when the train sample is finite or the tensor dimension $d = JK$ is large. We first shrink only the off-diagonal entries,

$$\widetilde{\Sigma}_\Delta = (1-\rho)\widehat{\Sigma}_\Delta + \rho\ \mathrm{diag}(\widehat{\Sigma}_\Delta), \tag{47}$$

which preserves featurewise variance while reducing noisy empirical correlations.

The parameter vector is a reshaped coefficient surface, so statistical stability alone is insufficient: neighboring coefficients should not oscillate without evidence. Let $\widetilde{P} \succeq 0$ be the normalized penalty matrix for second differences along both the log-probability and entropy axes, as in penalized B-spline smoothing (Eilers & Marx, 1996). With $s = \mathrm{tr}(\widetilde{\Sigma}_\Delta)/d$, define

$$A = \widetilde{\Sigma}_\Delta + s(\lambda\widetilde{P} + \gamma I). \tag{48}$$

Thus covariance shrinkage stabilizes correlations, $\lambda\widetilde{P}$ discourages an unnecessarily rough witness surface, and the ridge γI makes the system well posed.

Replacing the population covariance in the definition of $R(\theta)$ by A gives the oriented empirical signal-to-noise ratio

$$\widehat{R}_A(\theta) = \frac{\theta^\top\widehat{\delta}}{\sqrt{\theta^\top A\theta}}, \qquad \theta^\top\widehat{\delta} > 0. \tag{49}$$

Since the ratio is nonnegative after orientation, maximizing it is equivalent to maximizing its square. This produces the empirical Rayleigh problem

$$\max_{\theta\neq 0}\ \mathcal{J}_T(\theta), \qquad \mathcal{J}_T(\theta) = \frac{(\theta^\top\widehat{\delta})^2}{\theta^\top A\theta}. \tag{50}$$

When $\gamma > 0$, A is positive definite and the solution direction is

$$\widetilde{\theta} = A^{-1}\widehat{\delta}. \tag{51}$$

The closed-form derivation below uses whitening and Cauchy–Schwarz.

Because (50), the ranking induced by (8), and its conformal ranks are invariant to positive rescaling of θ, let $\overline{V}$ denote the train-set average of V_X and choose the following train-only numerical scale convention:

$$\theta_0 = \frac{\widetilde{\theta}}{\sqrt{\widetilde{\theta}^\top\overline{V}\widetilde{\theta}}} \tag{52}$$

when the denominator is nonzero. We orient its sign so that the train machine mean exceeds the train human mean. Spline dimensions and (ρ, λ, γ) are selected by train-only cross-validation; the selected direction is then refit once and frozen before calibration or testing.

Closed-form direction. Because $\widetilde{\Sigma}_\Delta \succeq 0$, $\widetilde{P} \succeq 0$, $s > 0$, and $\gamma > 0$, for every nonzero x,

$$x^\top Ax \geq s\gamma\|x\|_2^2 > 0.$$

Thus $A \succ 0$. Let $a = A^{1/2}\theta$ and $b = A^{-1/2}\widehat{\delta}$. Then

$$\mathcal{J}_T(\theta) = \frac{(a^\top b)^2}{a^\top a} \leq b^\top b = \widehat{\delta}^\top A^{-1}\widehat{\delta}.$$

Equality holds iff $a \propto b$, equivalently $\theta \propto A^{-1}\widehat{\delta}$. The condition $\widehat{\delta} \neq 0$ ensures that the optimum is nonzero and the quotient is not identically zero.

B ALGORITHM AND EXPERIMENTAL SETTINGS

Algorithm 1 Entropy-conditioned data-adaptive detection

Input: paired train sets $\mathcal{D}_{\mathrm{tr}}^{H}, \mathcal{D}_{\mathrm{tr}}^{M}$, human calibration set $\mathcal{C} = \{C_i\}_{i=1}^{m}$, query x, models q, r, frozen $\widehat{F}_H, B, C$, selected train-only hyperparameters, level α
Output: score s_x, conformal value $\widehat{p}(x)$, decision $\widehat{\varphi}_\alpha(x)$

function DOCUMENTFEATURE(X)
  $U \leftarrow \mathbf{0}_{JK}$
  for $t = 1, \ldots, |X|$ **do**
    $q_t(v) \leftarrow q(v \mid X_{<t}), \quad z_t(v) \leftarrow \log r(v \mid X_{<t})$
    $u_t \leftarrow \widehat{F}_H(-\sum_v q_t(v) \log q_t(v))$
    $\phi_t(v) \leftarrow B(z_t(v)) \otimes C(u_t), \quad \mu_t \leftarrow \sum_v q_t(v)\phi_t(v)$
    $U \leftarrow U + \phi_t(X_t) - \mu_t$
  end for
  return U
end function
for $X \in \mathcal{D}_{\mathrm{tr}}^{H} \cup \mathcal{D}_{\mathrm{tr}}^{M}$ **do**
  $U_X \leftarrow$ DOCUMENTFEATURE(X)
end for ▷ compute train features
for $c \in \{H, M\}$ **do**
  $\overline{U}_c \leftarrow n_c^{-1} \sum_{X \in \mathcal{D}_{\mathrm{tr}}^{c}} U_X$
end for
$\widehat{\delta} \leftarrow \overline{U}_M - \overline{U}_H$
$A \leftarrow$ regularized train covariance from Appendix A.7
$\widetilde{\theta} \leftarrow A^{-1}\widehat{\delta}$ ▷ learn witness direction
Orient $\widehat{\theta}$ so that $\widehat{\theta}^\top \widehat{\delta} > 0$
for $X \in \mathcal{C} \cup \{x\}$ **do**
  $U_X \leftarrow$ DOCUMENTFEATURE(X)
  $s_X \leftarrow T_{\widehat{\theta}}(X) = \widehat{\theta}^\top U_X$
end for ▷ score calibration and query texts
$\widehat{p}(x) \leftarrow [1 + \sum_{i=1}^{m} \mathbf{1}\{s_{C_i} \geq s_x\}]/(m+1)$
$\widehat{\varphi}_\alpha(x) \leftarrow \mathbf{1}\{\widehat{p}(x) \leq \alpha\}$ ▷ calibrated decision
return $s_x, \widehat{p}(x), \widehat{\varphi}_\alpha(x)$

Degree-two log-probability B-splines use raw scores, upper endpoint zero, and training-human lower/inner knots; entropy is mapped by the training-human empirical CDF and clipped to $[0.005, 0.995]$. Candidate tensor sizes are $4{\times}4$, $6{\times}6$, and $8{\times}6$, with $\lambda \in \{0, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, $\gamma \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, and $\rho \in \{0, 0.05, 0.2\}$. For second-difference operators D_z, D_h, we use

$$\widetilde{P} = \left[\frac{D_z^\top D_z \otimes I_K}{(J-2)K} + \frac{I_J \otimes D_h^\top D_h}{J(K-2)} \right] \Big/ \left(r^{-1} \sum_{\ell=1}^{r} \xi_\ell \right),$$

where $\xi_1, \ldots, \xi_r$ are the raw matrix's positive eigenvalues (unit scale if $r = 0$), making λ comparable across tensor sizes.

The symmetric eigensolver floors system eigenvalues at $\max\{\varepsilon s, \epsilon_{\mathrm{mach}} \max(\lambda_{\max}(A), 1)\}$; degenerate variance normalization uses Euclidean normalization, and $\mathrm{sign}(0) = 1$. Candidate tensor dimensions and (λ, γ, ρ) are selected by paired five-fold train-only cross-validation. The winning configuration is refit on all train passages, oriented using train class means, and then frozen before any calibration or test score is computed.

At deployment, Ours computes only the centered feature sum U_X and returns $T_{\widehat{\theta}}(X) = \widehat{\theta}^\top U_X$. The passage-specific covariance V_X is not required for primary query scoring; it is retained in the implementation only for train-time scale normalization and the locked $S_{\widehat{\theta}}$ deployment ablation.

For the conditional-curvature white-box comparison, Ours and official analytic AdaDetectGPT (Zhou et al., 2025) compute conditional moments exactly over the source-model vocabulary. Official Fast-DetectGPT (Bao et al., 2024) uses 10,000 conditional candidates. Only these official/full configurations are reported. In the proxy-model black-box evaluation, Fast, Ada, and Ours use exact q_t-weighted moments over the shared GPT-J vocabulary; GPT-Neo supplies the scoring

probabilities. The black-box Full ablation uses the same held-out-dataset protocol and exact proxy moments as the main Ours row, and the two result paths agree numerically in all 20 configurations.

The frozen split seed is 20260729. Within each source-model–dataset configuration, eligible human documents receive a deterministic random priority and are selected without replacement after excluding train indices, normalized text duplicates, and 48-word prefix collisions. Machine passages use a 120-token source prefix, top-$k = 50$ sampling, temperature 1, top-$p = 1$, minimum and maximum sequence lengths of 150 and 200 tokens, and at least 55 generated words. These decoding values are supplied explicitly rather than inherited from a checkpoint's generation configuration. Human/machine pairs are trimmed to their shorter word count. Each configuration contains 500 paired train passages, 500 disjoint human calibration passages, and 500 paired independent-test passages. Section 5 gives the evaluation order and reports every operating-point convention.

C Independent Human Calibration and Type-I Control

After the entropy transform, spline family, hyperparameters, score orientation, and $\widehat{\theta}$ have all been frozen using the train set, let $C_1, \ldots, C_m$ be an independent human calibration set. For a query x, define the upper-tail conformal rank p-value (Vovk et al., 2005)

$$\widehat{p}(x) = \frac{1 + \sum_{i=1}^{m} \mathbf{1}\{T_{\widehat{\theta}}(C_i) \geq T_{\widehat{\theta}}(x)\}}{m+1}, \tag{53}$$

and the level-α decision

$$\widehat{\varphi}_\alpha(x) = \mathbf{1}\{\widehat{p}(x) \leq \alpha\}. \tag{54}$$

AUROC, partial AUC, and PR-AUC use the continuous primary score $T_{\widehat{\theta}}(x)$; the fixed-FPR decision uses (53)–(54).

Theorem 4 (Finite-sample human type-I control) *After the score is frozen, suppose the calibration passages and a future human passage are exchangeable. Then the rule based on* (53) *satisfies*

$$\mathbb{P}\{\widehat{p}(X) \leq \alpha\} \leq \alpha$$

for every m and every score distribution.

Proof. Condition on the frozen score function and write $T_i = T_{\widehat{\theta}}(C_i)$ for $i \leq m$ and $T_{m+1} = T_{\widehat{\theta}}(X)$. Exchangeability makes the descending rank of T_{m+1} among these $m+1$ scores uniform in the absence of ties. Moreover,

$$\widehat{p}(X) = \frac{\#\{i \in \{1, \ldots, m+1\} : T_i \geq T_{m+1}\}}{m+1}.$$

With ties this quantity is at least the randomly tie-broken descending rank divided by $m+1$, so it is conservative. Consequently,

$$\mathbb{P}\{\widehat{p}(X) \leq k/(m+1)\} \leq k/(m+1)$$

for each integer k. Taking $k = \lfloor \alpha(m+1) \rfloor$ proves the claim.

The guarantee is marginal over a future human passage and the calibration set. It does not provide conditional control for every dataset, topic, or author subgroup, and it can fail if calibration data are reused for fitting, orientation, or hyperparameter/model selection. Conversely, the Gaussian, shared-covariance, and combined-MCLT conditions are irrelevant to this rank proof: their failure may reduce power, but it does not invalidate type-I control when the frozen score and exchangeability conditions hold.

D Complete Experimental Results

The complete tables below report the primary score $T_{\widehat{\theta}}$. A passage-normalized alternative is introduced only inside the explicitly labeled deployment ablation.

Table 3: Complete white-box AUROC underlying Table 1. Fast-DetectGPT uses the official 10,000-candidate configuration and AdaDetectGPT uses the official analytic configuration. All rows use the completed independent full-text evaluation; LLaMA-2-7B uses the corrected passage set.

Dataset	Method	Source Model				Avg.
		LLaMA-2-7B	OPT-2.7B	OPT-6.7B	Falcon-7B	
C4	Likelihood	0.5241	0.6047	0.5910	0.5631	0.5707
	Entropy	0.5614	0.5461	0.5419	0.5507	0.5500
	Rank	0.5767	0.6069	0.5979	0.5901	0.5929
	LogRank	0.5459	0.6317	0.6137	0.5892	0.5951
	LRR	0.6130	0.6881	0.6618	0.6584	0.6553
	Fast-DetectGPT	0.7147	0.8047	0.7904	0.7626	0.7681
	AdaDetectGPT	0.7903	0.8486	0.8334	0.8287	0.8253
	Ours	**0.7959**	**0.8784**	**0.8539**	**0.8509**	**0.8447**
PubMed	Likelihood	0.5255	0.6156	0.6078	0.5420	0.5727
	Entropy	0.5572	0.5593	0.5603	0.5532	0.5575
	Rank	0.5707	0.6124	0.6058	0.5787	0.5919
	LogRank	0.5400	0.6531	0.6428	0.5695	0.6014
	LRR	0.5781	0.7277	0.7089	0.6391	0.6635
	Fast-DetectGPT	0.6803	0.8099	0.7983	0.7123	0.7502
	AdaDetectGPT	0.7183	0.8583	0.8466	0.7717	0.7987
	Ours	**0.7221**	**0.8676**	**0.8605**	**0.7845**	**0.8087**
SQuAD	Likelihood	0.5143	0.6214	0.6042	0.5681	0.5770
	Entropy	0.5455	0.5411	0.5523	0.5506	0.5474
	Rank	0.5603	0.6014	0.5933	0.5938	0.5872
	LogRank	0.5306	0.6468	0.6269	0.5894	0.5984
	LRR	0.5806	0.6934	0.6706	0.6446	0.6473
	Fast-DetectGPT	0.6470	0.8147	0.8031	0.7716	0.7591
	AdaDetectGPT	0.7502	0.8505	0.8353	0.8270	0.8157
	Ours	**0.7516**	**0.8637**	**0.8522**	**0.8370**	**0.8261**
Writing	Likelihood	0.6292	0.7076	0.6976	0.6511	0.6714
	Entropy	0.5597	0.5164	0.5205	0.5246	0.5303
	Rank	0.6315	0.6218	0.6200	0.6037	0.6193
	LogRank	0.6530	0.7238	0.7132	0.6734	0.6909
	LRR	0.6833	0.7360	0.7271	0.7071	0.7134
	Fast-DetectGPT	0.8353	0.8791	0.8763	0.8451	0.8589
	AdaDetectGPT	0.8886	0.9111	0.9129	0.8979	0.9026
	Ours	**0.9002**	**0.9245**	**0.9220**	**0.9012**	**0.9120**
XSum	Likelihood	0.4981	0.5976	0.5795	0.5397	0.5537
	Entropy	0.5982	0.5566	0.5627	0.5870	0.5761
	Rank	0.5644	0.5920	0.5804	0.5740	0.5777
	LogRank	0.5207	0.6209	0.6032	0.5669	0.5779
	LRR	0.5809	0.6437	0.6352	0.6263	0.6215
	Fast-DetectGPT	0.6825	0.7647	0.7421	0.7262	0.7289
	AdaDetectGPT	0.7462	0.8062	0.7888	0.7777	0.7797
	Ours	**0.7571**	**0.8244**	**0.8070**	**0.7941**	**0.7956**

Table 4: Complete black-box AUROC underlying Table 2. GPT-J-6B and GPT-Neo-2.7B are the sampling/scoring proxies for the three conditional-curvature methods, while the probability/rank baselines use GPT-Neo-2.7B alone. All rows use the completed independent full-text evaluation; LLaMA-2-7B uses the corrected passage set.

Dataset	Method	Source Model				Avg.
		LLaMA-2-7B	OPT-2.7B	OPT-6.7B	Falcon-7B	
C4	Likelihood	0.5099	0.5702	0.5750	0.5421	0.5493
	Entropy	0.5336	0.5381	0.5218	0.5245	0.5295
	Rank	0.5334	0.5891	0.5829	0.5708	0.5690
	LogRank	0.5268	0.5938	0.5955	0.5605	0.5692
	LRR	0.5652	0.6437	0.6387	0.6056	0.6133
	Fast-DetectGPT	0.6200	0.7665	0.7427	0.6846	0.7034
	AdaDetectGPT	0.6930	0.8169	0.7905	0.7486	0.7623
	Ours	**0.7080**	**0.8363**	**0.8062**	**0.7599**	**0.7776**
PubMed	Likelihood	0.4660	0.3770	0.4224	0.4752	0.4351
	Entropy	0.5417	0.6226	0.5958	0.5464	0.5766
	Rank	0.5176	0.5360	0.5528	0.5509	0.5393
	LogRank	0.4763	0.4009	0.4472	0.4948	0.4548
	LRR	0.5217	0.5152	0.5452	0.5672	0.5373
	Fast-DetectGPT	0.5497	0.5642	0.6016	0.5899	0.5763
	AdaDetectGPT	0.5984	0.6563	0.6695	0.6724	0.6492
	Ours	**0.6351**	**0.7640**	**0.7636**	**0.7077**	**0.7176**
SQuAD	Likelihood	0.5059	0.5383	0.5536	0.5290	0.5317
	Entropy	0.5253	0.5466	0.5422	0.5326	0.5367
	Rank	0.5266	0.5842	0.5867	0.5765	0.5685
	LogRank	0.5182	0.5639	0.5755	0.5484	0.5515
	LRR	0.5536	0.6297	0.6309	0.5992	0.6034
	Fast-DetectGPT	0.6053	0.7483	0.7509	0.6712	0.6939
	AdaDetectGPT	0.6595	0.7989	0.7894	0.7298	0.7444
	Ours	**0.6876**	**0.8356**	**0.8237**	**0.7621**	**0.7773**

Continued on next page

Table 4 (continued)

Dataset	Method	Source Model				Avg.
		LLaMA-2-7B	OPT-2.7B	OPT-6.7B	Falcon-7B	
Writing	Likelihood	0.6121	0.6590	0.6705	0.6307	0.6431
	Entropy	0.5071	0.5087	0.4944	0.4932	0.5008
	Rank	0.5870	0.6155	0.6131	0.6050	0.6052
	LogRank	0.6206	0.6746	0.6842	0.6462	0.6564
	LRR	0.6248	0.6906	0.6950	0.6646	0.6687
	Fast-DetectGPT	0.7565	0.8372	0.8346	0.7690	0.7993
	AdaDetectGPT	**0.7973**	0.8756	**0.8771**	**0.8175**	**0.8419**
	Ours	0.7846	**0.8848**	0.8760	0.8161	0.8404
XSum	Likelihood	0.4662	0.5374	0.5351	0.4983	0.5093
	Entropy	0.5653	0.5498	0.5433	0.5551	0.5534
	Rank	0.5104	0.5652	0.5568	0.5553	0.5469
	LogRank	0.4789	0.5573	0.5559	0.5143	0.5266
	LRR	0.5127	0.5879	0.5931	0.5532	0.5617
	Fast-DetectGPT	0.5778	0.7000	0.6916	0.6213	0.6477
	AdaDetectGPT	0.6447	0.7600	0.7506	0.6963	0.7129
	Ours	**0.6691**	**0.7871**	**0.7792**	**0.7152**	**0.7377**

D.1 Operating-Point Results

Table 5: Operating-point metrics for the primary linear score, macro averaged over all 20 configurations. pAUC@5% uses the recorded standardized convention; FPR and power use independent human conformal calibration.

Setting	pAUC@5%	FPR@1%	Power@1%	FPR@5%	Power@5%
White-box	0.6424	0.0084	0.1772	0.0440	0.4002
Black-box	0.1807	0.0105	0.0999	0.0460	0.2714

Table 6: White-box operating-point metrics averaged over five datasets for each source model. LLaMA-2-7B uses the corrected passage set. Fast-DetectGPT uses the official 10,000-candidate configuration, AdaDetectGPT uses the official exact configuration, and Ours uses $T_{\widehat{\theta}}$.

Source Model	Method	pAUC@5%	FPR@1%	Power@1%	FPR@5%	Power@5%
LLaMA-2-7B	Fast-DetectGPT	0.5458	0.0056	0.0516	0.0540	0.2000
	AdaDetectGPT	0.5866	0.0076	0.0864	0.0532	0.3008
	Ours	**0.6031**	0.0088	**0.1252**	0.0460	**0.3104**
OPT-2.7B	Fast-DetectGPT	0.5962	0.0056	0.1144	0.0512	0.3400
	AdaDetectGPT	0.6455	0.0064	0.1712	0.0456	0.4252
	Ours	**0.6760**	0.0084	**0.2348**	0.0468	**0.4772**
OPT-6.7B	Fast-DetectGPT	0.5884	0.0088	0.1072	0.0452	0.2916
	AdaDetectGPT	0.6269	0.0080	0.1336	0.0420	0.3720
	Ours	**0.6525**	0.0096	**0.1984**	0.0376	**0.4184**
Falcon-7B	Fast-DetectGPT	0.5698	0.0092	0.0792	0.0432	0.2532
	AdaDetectGPT	0.6141	0.0084	0.1260	0.0404	0.3388
	Ours	**0.6379**	0.0068	**0.1504**	0.0456	**0.3948**

D.2 Class-Conditional Normality Tests

For each frozen white-box score distribution, we use a fitted-normal Kolmogorov–Smirnov statistic of the Lilliefors type (Lilliefors, 1967). Location and scale are re-estimated for every one of 999 Monte Carlo samples, and the resulting 40 classwise p-values are adjusted together by the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995). Table 7 reports the adjusted values. At the 5% level, none of the 20 machine-score distributions is rejected, whereas 9 of the 20 human-score distributions are rejected. The approximation is therefore treated as a fitting model rather than an assertion of exact finite-passage normality. Operational FPR control continues to come from the independent human calibration set.

Table 7: BH-adjusted q-values for fitted-normal Monte Carlo tests. Each entry is human/machine; bold values reject normality at 5%.

Source model	C4	PubMed	SQuAD	Writing	XSum
Falcon-7B	0.056/0.497	**0.031**/0.120	**0.031**/0.497	**0.008**/0.824	0.405/0.303
LLaMA-2-7B	**0.008**/0.497	**0.008**/0.491	0.065/0.282	**0.031**/0.480	0.235/0.480
OPT-2.7B	0.133/0.720	0.531/0.497	0.531/0.824	0.312/0.339	**0.031**/0.198
OPT-6.7B	**0.008**/0.824	0.478/0.480	0.120/0.405	**0.008**/0.491	0.120/0.824

D.3 Ablation Results

Passage-specific normalization. For this ablation only, define

$$S_\theta(X) = \frac{T_\theta(X)}{\sqrt{\max\{\theta^\top V_X \theta, \varepsilon\}}}. \tag{55}$$

The primary detector deploys T_θ; this alternative divides each passage by a different reference-law variance factor.

Table 8: Change from S_θ to T_θ. Every entry is $T - S$, macro averaged over 20 configurations.

Setting	Split	AUROC	PR-AUC	pAUC@5%	Power@1%	Power@5%
White	Independent	+0.0034	+0.0076	+0.0123	+0.0366	+0.0239
White	Fresh	+0.0045	+0.0106	+0.0257	+0.0308	+0.0105
Black	Independent	+0.0033	+0.0062	+0.0111	+0.0181	+0.0242
Black	Fresh	+0.0038	+0.0077	+0.0111	+0.0201	+0.0229

On the independent split, the denominator coefficient of variation averages 11.5% white-box and 10.2% black-box, so it is not a constant rescaling. Machine passages have a larger mean denominator in 17/20 and 18/20 configurations, respectively. The machine–human denominator difference correlates 0.705 white-box and 0.847 black-box with the AUROC gain from removing it. These are descriptive diagnostics, not assumptions used to fit the score.

Component factorials under the primary score. We next evaluate the same 2^3 design in both access regimes using T_θ. The switches H, S, and C denote entropy conditioning, surface-smoothness selection, and covariance-shrinkage selection, respectively. Each entry is macro averaged over the same 20 source-model–dataset configurations. The FPR/Power columns report realized test FPR and machine-text power after independent human calibration at the stated nominal level.

In the white-box factorial, every non-Full variant selects its permitted hyperparameters by paired train-only cross-validation and is then refit before calibration and testing. The Full row uses the frozen primary fit and therefore reproduces the main white-box result.

Table 9: White-box component factorial under the primary score T_θ on the independent test, macro averaged over 20 configurations.

Variant	H	S	C	AUROC	pAUC@5%	FPR@1%	Pow.@1%	FPR@5%	Pow.@5%
B_0	0	0	0	0.8259	0.2750	0.0076	0.1596	0.0442	0.3787
$B_0 + H$	1	0	0	0.8367	0.2998	0.0087	0.1772	0.0413	0.3986
$B_0 + S$	0	1	0	0.8258	0.2748	0.0078	0.1602	0.0446	0.3787
$B_0 + C$	0	0	1	0.8255	0.2747	0.0074	0.1582	0.0450	0.3770
$B_0 + H + S$	1	1	0	0.8372	**0.3029**	0.0084	**0.1782**	0.0437	0.3979
$B_0 + H + C$	1	0	1	0.8370	0.3009	0.0087	0.1746	0.0427	0.3986
$B_0 + S + C$	0	1	1	0.8255	0.2748	0.0072	0.1575	0.0453	0.3784
Full	1	1	1	**0.8374**	0.3026	0.0084	0.1772	0.0440	**0.4002**

For the black-box factorial, we freeze the directions and hyperparameters from the completed train-only factorial fits and recompute all calibration and test scores with T_θ. This cache-only rescore changes the deployment statistic without refitting a variant or rerunning either proxy model.

Table 10: Black-box component factorial under the primary score T_θ on the independent test, macro averaged over 20 configurations. Directions and hyperparameters are frozen from the corresponding train-only factorial fits.

Variant	H	S	C	AUROC	pAUC@5%	FPR@1%	Pow.@1%	FPR@5%	Pow.@5%
B_0	0	0	0	0.7459	0.1560	0.0140	0.0973	0.0519	0.2484
$B_0 + H$	1	0	0	0.7693	**0.1813**	0.0106	0.1009	0.0464	0.2695
$B_0 + S$	0	1	0	0.7460	0.1558	0.0136	0.0976	0.0511	0.2462
$B_0 + C$	0	0	1	0.7459	0.1560	0.0140	0.0973	0.0519	0.2484
$B_0 + H + S$	1	1	0	0.7692	0.1809	0.0105	**0.1016**	0.0475	**0.2731**
$B_0 + H + C$	1	0	1	0.7701	0.1804	0.0107	0.0997	0.0455	0.2698
$B_0 + S + C$	0	1	1	0.7460	0.1558	0.0136	0.0976	0.0511	0.2462
Full	1	1	1	**0.7701**	0.1807	0.0105	0.0999	0.0460	0.2714

Entropy conditioning gives the dominant AUROC main effect in both regimes: 0.01144 (95% CI $[0.00815, 0.01469]$) in white-box evaluation and 0.02368 ($[0.01596, 0.03280]$) in black-box evaluation. Its pAUC@5% effects are 0.02673 and 0.02493, respectively. In contrast, every AUROC main effect of S or C has magnitude below 0.0005. Entropy conditioning also improves calibrated power at nominal 5% FPR by 0.02063 white-box and 0.02365 black-box. At nominal 1% FPR, however, the corresponding confidence intervals include zero in both regimes, so the extreme-tail power effect is inconclusive.