AKTD: FDR Control for LLM Training Data Detection under Approximate Exchangeability

Jingcheng Wang[1] **Yingjie Wang**[2] **Yisheng Lv**[1,*]

[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]China University of Petroleum (East China), Qingdao, China

jingcheng.wang@ia.ac.cn yingjiewang@upc.edu.cn
yisheng.lv@ia.ac.cn

[*]Corresponding author

Abstract

Detecting whether specific text samples were used to train large language models is increasingly important for privacy, copyright, and data auditing. Knockoff-based training data detection (KTD) offers a promising approach to controlling the false discovery rate (FDR). However, its effectiveness relies on null sign-flip symmetry induced by exchangeability between each null sample and its knockoff. This condition is difficult to satisfy in natural language settings. We study training data detection under approximate exchangeability, and demonstrate that deviations from exact exchangeability induce measurable null-tail asymmetry, which in turn leads to systematic FDR inflation under the standard knockoff+ threshold. This effect can be captured by pathwise bounds, which explain when and why KTD becomes anti-conservative. Based on this, we propose Asymmetry-adjusted KTD (AKTD), a series of procedures that correct for asymmetry by estimating the penalty term. We introduce an external null variant that uses a separate calibration pool containing only null samples to estimate the penalty term. This enables detection with FDR control without using labels from the evaluated candidate set. Experiments on WikiMIA and MIMIR using GPT and Pythia demonstrate that the asymmetry term accurately predicts the observed FDR expansion, and AKTD restores FDR control while maintaining effective detection capabilities.

1 Introduction

Large Language Models (LLMs) are trained on massive corpora of text from the web, books, code repositories, and other sources. This scale has raised growing concerns about privacy, copyright, data governance, and benchmark contamination [Maini et al., 2024]. In many auditing scenarios, it is essential to determine whether a set of candidate texts was used to train a deployed model. This task is known as training data detection or member inference for LLMs, which is important in legal, scientific, and data governance settings. Reliable detection therefore requires more than high classification accuracy. It also requires control of the false discovery rate (FDR), which is the expected proportion of non-training samples among all reported members.

Knockoff-based training data detection (KTD) provides a promising approach to control the FDR by modeling training data detection as a multiple testing problem [Hu et al., 2025]. For each candidate text, KTD generates a knockoff text and compares the target model's responses to the original and knockoff samples. Under knockoff exchangeability, the negative tail of the obtained statistic can be used to estimate the FDR in the positive tail, allowing FDR controlled selection via the knockoff+ threshold.

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

In practice, existing text knockoff generators rarely satisfy exact exchangeability. Text imitations generated by paraphrasing or language models may retain semantic content but alter fluency, length, word choice, or likelihood under the target model.

These imitations break the null sign symmetry. For non-training samples, positive statistics may then appear more often than negative statistics of the same magnitude. In this case, the ordinary knockoff+ numerator can underestimate the number of false discoveries, and the KTD threshold may become insufficiently conservative.

We study KTD under approximate exchangeability. Our key observation is that the failure of exact exchangeability can be quantified by a realized null-tail asymmetry. This asymmetry is the excess of the positive tail over the negative tail among null statistics. We derive a pathwise bound on the false discovery proportion (FDP). Taking expectation gives an FDR bound with an explicit asymmetry penalty. The bound shows when KTD remains reliable and when it becomes anti-conservative.

Based on this analysis, we propose an Asymmetry-adjusted KTD (AKTD) framework, which corrects the knockoff+ threshold by adding an estimated asymmetry penalty. We study oracle AKTD for mechanism validation, cross-fit AKTD for label-assisted diagnostics, and external-null AKTD, the latter estimating the penalty from a single calibration pool containing only null values instead of labels from an evaluated candidate set.

Experiments conducted on WikiMIA and MIMIR using GPT-2 and Pythia-410M/1B/2.8B support this view. The proposed asymmetric term explains the empirical FDP inflation of ordinary KTD, while AKTD restores controlled selection in oracle and cross-fit diagnostics. When the external null distribution matches sufficiently, the external-null AKTD provides a calibration method without candidate set labels. Stress tests further show that the same-null assumption is substantive.

2 Related Work

Training data detection and membership inference. Training data detection is related to memorization, privacy leakage, copyright auditing, dataset-level inference, and benchmark contamination [Kandpal et al., 2022, Carlini et al., 2022b, Chang et al., 2023]. Membership inference for LLMs can be affected by distribution shift between member and non-member samples, while dataset-level auditing studies whether a whole dataset was used for training [Duan et al., 2024, Maini et al., 2024]. Most methods assign each text a membership score and evaluate accuracy, AUC, or TPR at a fixed FPR. Common scores include loss or likelihood and reference model comparisons [Carlini et al., 2022a], Zlib-normalized likelihood [Carlini et al., 2021], MinK% variants [Shi et al., 2023, Zhang et al., 2025], relative conditional log-likelihood [Xie et al., 2024], and perturbation-based scores [Mattern et al., 2023]. These scores can rank samples well, but they do not control the false-positive proportion among reported members. We instead view sample-level training data detection as a controlled selection problem, where the goal is to select likely members while controlling the expected fraction of false discoveries.

Knockoffs and approximate exchangeability. Knockoff inference controls FDR by comparing original variables with synthetic knockoffs. It uses the negative tail of a statistic with sign symmetry to estimate false discoveries in the positive tail [Barber and Candes, 2015, Candes et al., 2018]. KTD applies this idea to LLM training data detection by generating text knockoffs and using the knockoff threshold [Hu et al., 2025]. Its guarantee requires exact pairwise exchangeability between each null text and its knockoff. This condition gives sign-flip symmetry of the null statistics. However, it is hard to satisfy in natural language. Paraphrased or model-generated knockoffs may preserve meaning but change length, fluency, word choice, compression behavior, or likelihood under the target model. These shifts can make the positive tail of null statistics heavier than the negative tail. As a result, the standard knockoff+ numerator, which uses the negative tail as a surrogate for false discoveries in the positive tail, can underestimate the number of false discoveries [Barber et al., 2020, Fan et al., 2025a,b]. Therefore, we study text-level knockoffs for LLM training data detection and show that realized null-tail asymmetry is the key failure mode addressed by AKTD. Conceptually, robust-knockoff analyses bound distributional or coupling errors, whereas AKTD directly corrects the realized positive-versus-negative null-tail imbalance; Appendix D gives a brief comparison.

3 Preliminaries

Training data detection. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be the candidate set, and let f_θ be the target language model. Each candidate sample is either a member or a non-member of the training data. We denote the member indices by $\mathcal{H}_1 \subseteq [n]$ and the non-member indices by $\mathcal{H}_0 = [n] \setminus \mathcal{H}_1$. A detection procedure outputs a set $\widehat{S} \subseteq [n]$, where $j \in \widehat{S}$ means that X_j is declared to be a training sample.

For a selected set $\widehat{S}$, the false discovery proportion and false discovery rate are

$$\mathrm{FDP} = \frac{|\widehat{S} \cap \mathcal{H}_0|}{|\widehat{S}| \vee 1}, \qquad \mathrm{FDR} = \mathbb{E}[\mathrm{FDP}].$$

The expectation is taken over the candidate set, knockoff generation, calibration samples if used, and any algorithmic randomness. The model f_θ is treated as fixed. We also report power,

$$\mathrm{Power} = \mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{H}_1|}{|\mathcal{H}_1| \vee 1}\right].$$

The goal is to control $\mathrm{FDR} \leq q$ for a target level $q \in (0, 1)$ while retaining high power.

KTD statistic. For each candidate text X_j, KTD generates m knockoff texts $\widetilde{X}_j^{(1)}, \ldots, \widetilde{X}_j^{(m)}$. Let $s(X; \theta)$ be a membership score, where larger values indicate stronger evidence of membership. We use

$$W_j = s(X_j; \theta) - \frac{1}{m} \sum_{\ell=1}^{m} s(\widetilde{X}_j^{(\ell)}; \theta).$$

A positive W_j means that the original text receives stronger membership evidence than its knockoff references. Our analysis is stated at the level of the induced statistic vector $W = (W_1, \ldots, W_n)$.

Knockoff+ threshold. For $t > 0$, define

$$R(t) = \sum_{j=1}^{n} \mathbf{1}\{W_j \geq t\}, \qquad N(t) = \sum_{j=1}^{n} \mathbf{1}\{W_j \leq -t\},$$

where $R(t)$ is the number of positive selections and $N(t)$ is the observable negative tail count. The null part of $N(t)$ is used to estimate false discoveries; negative non-null statistics only make this count more conservative. Let

$$\mathcal{T} = \{|W_j| : |W_j| > 0,\ j \in [n]\}.$$

The ordinary KTD threshold is

$$\tau_{\mathrm{KTD}} = \inf\left\{t \in \mathcal{T} : \frac{1 + N(t)}{R(t) \vee 1} \leq q\right\},$$

with $\tau_{\mathrm{KTD}} = \infty$ if the set is empty. The selected set is

$$\widehat{S}_{\mathrm{KTD}} = \{j : W_j \geq \tau_{\mathrm{KTD}}\}.$$

When a threshold is ∞, the selected set is empty. All tail counts and asymmetry quantities introduced later are taken to be zero.

Null sign symmetry. Let $A_j = |W_j|$ and $S_j = \mathrm{sign}(W_j)$ for $W_j \neq 0$. Ordinary knockoff+ relies on null sign symmetry at the level of the test statistic. Given the magnitudes $\{|W_j| : j \in [n]\}$ and the non-null statistics $\{W_j : j \in \mathcal{H}_1\}$, the null signs $\{S_j : j \in \mathcal{H}_0,\ W_j \neq 0\}$ should be independent Rademacher variables. Under this condition, the negative null tail is a valid surrogate for the positive null tail. For statistics based on a single knockoff, exact pairwise exchangeability and antisymmetry imply this sign symmetry. For averaged statistics based on multiple knockoffs, the implication may fail when $m > 1$. We therefore analyze violations directly through realized null-tail asymmetry.

External-null calibration setup. For external-null AKTD, we use an independent calibration pool that contains only null samples. We process this pool with the same scoring and knockoff generation pipeline used for the evaluated candidate set. The certificate requires two conditions. The calibration statistics and the evaluated null statistics must follow a shared null distribution. In addition, we need a valid upper bound on the number of null candidates.

4 Methodology

We introduce AKTD, an asymmetry-aware framework for training data detection under approximate knockoff exchangeability. AKTD characterizes deviations from statistic-level null sign symmetry using realized null-tail asymmetry. It then incorporates this quantity into the selection threshold through an estimated asymmetry penalty.

4.1 Pathwise Diagnosis of Null-tail Asymmetry

For a threshold $t > 0$, define the positive and negative null tails

$$V_+(t) = \sum_{j \in \mathcal{H}_0} \mathbf{1}\{W_j \geq t\}, \qquad V_-(t) = \sum_{j \in \mathcal{H}_0} \mathbf{1}\{W_j \leq -t\}.$$

Here $V_+(t)$ is the false-discovery count at threshold t, while $V_-(t)$ is the negative null tail. Both are not observed in deployment because the null labels are unknown. Under statistic null sign symmetry, the standard knockoff+ argument gives

$$\mathbb{E}\left[\frac{V_+(\tau_{\mathrm{KTD}})}{1 + V_-(\tau_{\mathrm{KTD}})}\right] \leq 1.$$

Since $V_-(t) \leq N(t)$ for every t, ordinary KTD run at level q satisfies

$$\mathrm{FDR}_{\mathrm{KTD}} = \mathbb{E}\left[\frac{V_+(\tau_{\mathrm{KTD}})}{R(\tau_{\mathrm{KTD}}) \vee 1}\right] \leq q.$$

This ideal guarantee follows from the standard sign-flip argument. We recall the connection between pairwise exchangeability and statistic-level sign symmetry in Appendix A.1, and the FDR-control argument in Appendices A.2 and A.3.

To quantify violations of this surrogate, define the count-scale null-tail asymmetry$D_0(t) = \{V_+(t) - V_-(t)\}_+$, and the normalized asymmetry

$$\Delta_0(t) = \frac{D_0(t)}{R(t) \vee 1}, \qquad \Delta_0^\star = \sup_{t \in \mathcal{T}} \Delta_0(t).$$

If $\mathcal{T}$ is empty, we use the convention $\Delta_0^\star = 0$. The quantity $D_0(t)$ measures how much the positive null tail exceeds the negative null tail at threshold t.

Proposition 1 (Pathwise FDP deviation) *For every realized statistic vector W, the ordinary KTD threshold satisfies*

$$\mathrm{FDP}_{\mathrm{KTD}}(W) \leq q + \Delta_0(\tau_{\mathrm{KTD}}) \leq q + \Delta_0^\star.$$

Consequently,

$$\mathrm{FDR}_{\mathrm{KTD}} \leq q + \mathbb{E}[\Delta_0^\star].$$

Proposition 1 is deterministic and does not require a distributional assumption, as shown in Appendix A.4. It should be viewed as a pathwise diagnostic decomposition. The excess FDP of ordinary KTD is controlled by the realized imbalance between the positive and negative null tails. When $D_0(t)$ is small along the selected path, ordinary KTD remains reliable. When $D_0(t)$ is large, the ordinary knockoff+ ratio can become anti-conservative. In the experiments, we report $q + \Delta_0(\tau_{\mathrm{KTD}})$ as a pathwise diagnostic and compare it with the empirical FDP observed on the labeled benchmark.

Algorithm 1 AKTD with a Generic Asymmetry Penalty

Require: Statistics $\{W_j\}_{j=1}^n$, internal level q_{int}, penalty function $B(t)$
1: Form $\mathcal{T} = \{|W_j| : |W_j| > 0,\ j \in [n]\}$
2: If $\mathcal{T} = \emptyset$, return $\emptyset$
3: For all $t \in \mathcal{T}$, compute $R(t) = \sum_{j=1}^n \mathbf{1}\{W_j \geq t\}$ and $N(t) = \sum_{j=1}^n \mathbf{1}\{W_j \leq -t\}$
4: Define $\mathcal{A}_B = \{t \in \mathcal{T} : (1 + N(t) + B(t))/(R(t) \vee 1) \leq q_{\mathrm{int}}\}$
5: If $\mathcal{A}_B = \emptyset$, return $\emptyset$
6: Set $\tau_B = \inf \mathcal{A}_B$
7: **return** $\widehat{S}_B = \{j : W_j \geq \tau_B\}$

4.2 Asymmetry-Adjusted Thresholding

The pathwise bound suggests correcting the ordinary numerator by an asymmetry penalty. Let $B(t) \geq 0$ be a count-scale penalty intended to upper bound

$$D_0(t) = \{V_+(t) - V_-(t)\}_+.$$

Given an internal selection level $q_{\mathrm{int}} \in (0, 1)$, the AKTD threshold is

$$\tau_B = \inf\left\{t \in \mathcal{T} : \frac{1 + N(t) + B(t)}{R(t) \vee 1} \leq q_{\mathrm{int}}\right\},$$

with $\tau_B = \infty$ if the set is empty. The selected set is

$$\widehat{S}_B = \{j : W_j \geq \tau_B\}.$$

Proposition 2 (Adjusted-threshold certificate) *Suppose that the penalty satisfies $B(t) \geq D_0(t)$ for all $t \in \mathcal{T}$. Then the adjusted procedure run at internal level q_{int} satisfies $\mathrm{FDP}_B(W) \leq q_{\mathrm{int}}$ for every realized statistic vector W.*

Indeed,

$$V_+(t) \leq V_-(t) + D_0(t) \leq V_-(t) + B(t) \leq N(t) + B(t),$$

and evaluating this inequality at $t = \tau_B$ gives the certificate. The formal proof is given in Appendix A.5. Thus AKTD reduces controlled selection to the construction of a valid upper envelope for the realized null tail imbalance. Proposition 2 does not by itself make AKTD deployable; deployability depends on how the penalty $B(t)$ is obtained.

$$B_{\mathrm{oracle}}(t) = D_0(t)$$

and is used only as a mechanism check because it requires true null labels. Cross-fit AKTD estimates $B(t)$ from held-out benchmark null folds and serves only as a label-assisted diagnostic. External-null AKTD estimates $B(t)$ from a separate null-only calibration pool, avoiding candidate-set labels during thresholding. Its certificate requires the external and evaluated null statistics to share the same distribution. All variants use Algorithm 1; ordinary KTD is recovered by setting $B(t) \equiv 0$ and $q_{\mathrm{int}} = q$. Details are in Appendices A.9 and B.

4.3 External-Null Calibration

Let $C_1, \ldots, C_M$ be statistics computed on an external null-only calibration pool using the same score and knockoff-generation pipeline as the candidate statistics W_j. Define the population null tail difference

$$d_0(t) = \mathbb{P}(C \geq t) - \mathbb{P}(C \leq -t), \qquad a_0(t) = \{d_0(t)\}_+,$$

and its empirical estimate

$$\widehat{d}_M(t) = \frac{1}{M}\sum_{i=1}^M \left[\mathbf{1}\{C_i \geq t\} - \mathbf{1}\{C_i \leq -t\}\right], \qquad \widehat{a}_M(t) = \{\widehat{d}_M(t)\}_+.$$

The certified external-null procedure uses a statistic-level same-null calibration condition. Conditional on the evaluated null set $\mathcal{H}_0$, the evaluated null statistics $\{W_j : j \in \mathcal{H}_0\}$ and the external calibration statistics $C_1, \ldots, C_M$ are assumed to be independent draws from the same-null statistic distribution. We also assume a valid upper bound $|\mathcal{H}_0| \leq \bar{n}_0$. The conservative choice is $\bar{n}_0 = n$. A smaller value is allowed only with external justification. If $\bar{n}_0$ is underestimated, the certificate is no longer valid.

For a confidence level $\alpha \in (0, 1)$, define

$$\eta_M(\alpha) = \sqrt{\frac{2\log(4/\alpha)}{M}}, \qquad \eta_n(\alpha) = \sqrt{2\bar{n}_0 \log(4/\alpha)}.$$

The certified external-null penalty is

$$B_\alpha(t) = \bar{n}_0\{\widehat{a}_M(t) + \eta_M(\alpha)\} + \eta_n(\alpha).$$

With high probability, this penalty provides a uniform upper bound on $D_0(t)$ over all thresholds t. Because the concentration bound is uniform, it also applies to the threshold selected by Algorithm 1. See Appendix A.6 for details.

The assumption of a shared null distribution is important in LLM auditing. The external null pool must remain representative after the same scoring and knockoff generation pipeline is applied. If the external null statistics are shifted relative to the evaluated null statistics, the penalty may underestimate $D_0(t)$. The certificate then no longer applies. Thus, the same-null condition is a substantive statistic-level transfer assumption, not merely a requirement that calibration texts be non-members; Appendix D summarizes its scope.

Theorem 1 (Certified external-null AKTD) *Under the same-null calibration assumption and the null-count bound $|\mathcal{H}_0| \leq \bar{n}_0$, the AKTD procedure with $B(t) = B_\alpha(t)$ and internal level q_{int} satisfies*

$$\mathbb{P}\left(\mathrm{FDP}_{\mathrm{ext}}(W) \leq q_{\mathrm{int}}\right) \geq 1 - \alpha.$$

Consequently,

$$\mathrm{FDR}_{\mathrm{ext}} \leq q_{\mathrm{int}}(1 - \alpha) + \alpha \leq q_{\mathrm{int}} + \alpha.$$

Theorem 1 gives a high-probability FDP certificate. It does not give exact FDR control at the same internal level. The proof is given in Appendix A.7. To control FDR at a target level q, one can run the certified procedure at

$$q_{\mathrm{int}} = q_0 = \frac{q - \alpha}{1 - \alpha}, \qquad 0 < \alpha < q < 1,$$

which yields $\mathrm{FDR} \leq q$. The simpler choice $q_{\mathrm{int}} = q - \alpha$ is also conservative when $\alpha \leq q$; see Appendix A.8. The certified penalty $B_\alpha(t)$ is conservative, but it comes with a finite-sample guarantee. In experiments, we also test bootstrap and quantile penalties for the same asymmetry on the count scale. These empirical penalties can improve power, but they are not guaranteed in finite samples and may fail when the external null distribution is mismatched. The formal guarantee applies to $B_\alpha(t)$, or to any penalty that satisfies the domination condition in Proposition 2.

5 Experiments

Our experiments answer three main questions. First, does ordinary KTD fail in the way predicted by the null-tail asymmetry analysis? Second, can AKTD correct this failure while keeping useful power? Third, how does AKTD relate to strong score baselines?

5.1 Experimental Setup

We evaluate GPT-2 and Pythia-410M/1B/2.8B on WikiMIA and MIMIR. WikiMIA has 861 members and 789 non-members. MIMIR-Arxiv has 500 members and 500 non-members, and MIMIR-PileCC has 1000 members and 1000 non-members. WikiMIA is used mainly for pathwise and oracle diagnostics. MIMIR is used for external-null calibration. All experiments were run on a single NVIDIA GeForce RTX 4090 GPU with 24GB memory.

Table 1: Pathwise asymmetry diagnosis at $q = 0.1$. Entries report empirical FDP/power; AKTD (O/C) uses oracle/cross-fit penalties. Bold FDP values meet the target.

Model	Dataset	Statistic	$q + \Delta_0$ (τ)	KTD Emp. FDP↓ / Power↑	**AKTD (O)** Emp. FDP↓ / Power↑	**AKTD (C)** Emp. FDP↓ / Power↑
GPT-2	WikiMIA	Grad. norm	0.204	0.201 / 0.956	**0.098** / 0.811	**0.092** / 0.734
GPT-2	MIMIR-Arxiv	Log-prob.	0.600	0.499 / 1.000	**0.095** / 0.724	**0.073** / 0.684
GPT-2	MIMIR-PileCC	Log-prob.	0.593	0.497 / 1.000	**0.096** / 0.590	**0.079** / 0.527
Pythia-410M	MIMIR-Arxiv	Log-prob.	0.600	0.499 / 1.000	**0.096** / 1.000	**0.076** / 1.000
Pythia-410M	MIMIR-PileCC	Log-prob.	0.598	0.499 / 1.000	**0.098** / 1.000	**0.084** / 1.000
Pythia-1B	MIMIR-Arxiv	Log-prob.	0.600	0.499 / 1.000	**0.096** / 1.000	**0.069** / 1.000
Pythia-1B	MIMIR-PileCC	Log-prob.	0.600	0.500 / 1.000	**0.098** / 1.000	**0.087** / 1.000
Pythia-2.8B	MIMIR-Arxiv	Log-prob.	0.600	0.499 / 1.000	**0.096** / 1.000	**0.081** / 1.000
Pythia-2.8B	MIMIR-PileCC	Log-prob.	0.599	0.499 / 1.000	**0.098** / 1.000	**0.088** / 1.000

We compare ordinary KTD with three AKTD variants. AKTD (O) uses the true null-tail asymmetry to verify the mechanism. AKTD (C) estimates the penalty from held-out benchmark null folds and is a label-assisted diagnostic protocol. AKTD (E) estimates the penalty from an external null-only calibration pool and does not use candidate-set labels during thresholding. Its validity requires the external and evaluated null statistics to be matched.

All practical AKTD results use a bootstrap-quantile asymmetry estimator with 300 bootstrap repetitions, quantile 0.8, smoothing window 1, and five cross-fitting folds. We also report LOSS, MinK%, and Zlib as score-only baselines. Their label-tuned operating points use benchmark labels to choose thresholds. They are therefore diagnostic upper envelopes, not deployable FDR-control procedures. Additional experimental details are provided in Appendix C.

5.2 Pathwise Asymmetry Diagnosis

Table 1 reports the main pathwise diagnostic results. Ordinary KTD can be severely anti-conservative when generated knockoffs break null sign symmetry. On WikiMIA/GPT-2, ordinary KTD has empirical FDP 0.201 at $q = 0.1$, and the pathwise bound $q + \Delta_0(\tau)$ is 0.204. Thus the bound closely tracks the observed inflation. On MIMIR, ordinary KTD selects nearly all hypotheses and has empirical FDP close to 0.5.

Pythia-410M/1B/2.8B on MIMIR follow the same pattern. Ordinary KTD has empirical FDP near one half with nearly full selection, while oracle and cross-fit AKTD stay below the $q = 0.1$ target without losing power. This consistency reinforces the mechanism evidence and does not indicate a scale specific anomaly. The pattern also holds for Pythia-6.9B, Pythia-12B, and GPT-J-6B: ordinary KTD remains anti-conservative, whereas oracle and cross-fit AKTD remain below the target with high power (Appendix D).

Oracle AKTD restores empirical FDP below the target in all settings. This supports the mechanism claim that null-tail asymmetry drives the failure of ordinary KTD. Cross-fit AKTD also keeps empirical FDP below $q = 0.1$, while using only held-out benchmark null folds to estimate the penalty. In addition, Appendix C.1 further reports the WikiMIA/GPT-2 diagnosis across multiple target levels. Appendix C.2 then motivates the default bootstrap-quantile asymmetry estimator, and Appendix C.3 motivates the log-probability statistic used for MIMIR after a statistic ablation.

These results show that exact sign symmetry is not a benign detail. When the positive null tail is larger than the negative null tail, the ordinary knockoff+ numerator $1 + N(t)$ underestimates false discoveries in the positive tail. AKTD corrects the threshold by adding an estimated penalty for this realized imbalance.

5.3 External-Null Calibration

We next evaluate AKTD (E), which is the candidate-set label-free variant. It estimates the asymmetry penalty from an external null-only calibration pool. For MIMIR, we use the other MIMIR domain's null subset as the external pool. This is a cross domain stress test rather than a perfectly matched same-null setting.

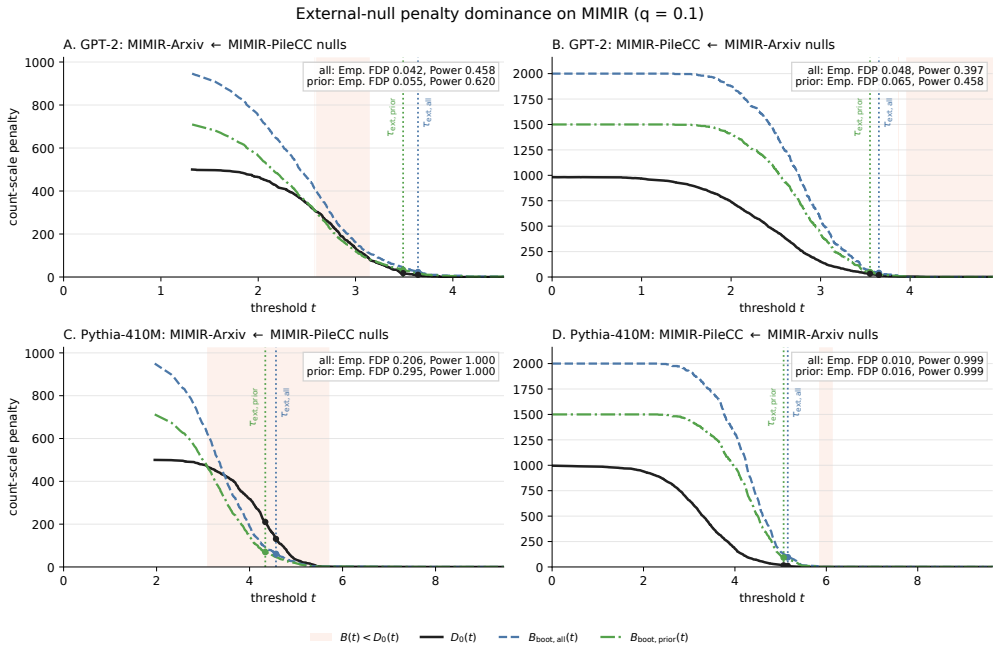

Figure 1: Penalty-dominance diagnostics for External-null AKTD on MIMIR at $q = 0.1$. Curves compare the realized asymmetry $D_0(t)$ with external bootstrap penalties; shaded regions mark thresholds where the penalty falls below $D_0(t)$, and dotted lines mark selected thresholds.

Table 2: External-null calibration at $q = 0.1$. Entries report empirical FDP/power; AKTD (E) estimates the penalty from external nulls without candidate-set labels. The n_0 bound is either all-null or a prior bound $n_0 \leq 0.75n$. Bold FDP values meet the target; † marks external-null mismatch.

Model	Dataset	External null pool	n_0 bound	KTD Emp. FDP↓ / Power↑	**AKTD (E)** Emp. FDP↓ / Power↑
GPT-2	MIMIR-Arxiv	MIMIR-PileCC	all	0.499 / 1.000	**0.042** / 0.458
GPT-2	MIMIR-Arxiv	MIMIR-PileCC	prior	0.499 / 1.000	**0.055** / 0.620
GPT-2	MIMIR-PileCC	MIMIR-Arxiv	all	0.497 / 1.000	**0.048** / 0.397
GPT-2	MIMIR-PileCC	MIMIR-Arxiv	prior	0.497 / 1.000	**0.065** / 0.458
Pythia-410M	MIMIR-Arxiv	MIMIR-PileCC	all	0.499 / 1.000	$0.206^{\dagger}$ / 1.000
Pythia-410M	MIMIR-Arxiv	MIMIR-PileCC	prior	0.499 / 1.000	$0.295^{\dagger}$ / 1.000
Pythia-410M	MIMIR-PileCC	MIMIR-Arxiv	all	0.499 / 1.000	**0.010** / 0.999
Pythia-410M	MIMIR-PileCC	MIMIR-Arxiv	prior	0.499 / 1.000	**0.016** / 0.999
Pythia-1B	MIMIR-Arxiv	MIMIR-PileCC	all	0.499 / 1.000	$0.237^{\dagger}$ / 1.000
Pythia-1B	MIMIR-Arxiv	MIMIR-PileCC	prior	0.499 / 1.000	$0.312^{\dagger}$ / 1.000
Pythia-1B	MIMIR-PileCC	MIMIR-Arxiv	all	0.500 / 1.000	**0.006** / 1.000
Pythia-1B	MIMIR-PileCC	MIMIR-Arxiv	prior	0.500 / 1.000	**0.007** / 1.000
Pythia-2.8B	MIMIR-Arxiv	MIMIR-PileCC	all	0.499 / 1.000	$0.182^{\dagger}$ / 1.000
Pythia-2.8B	MIMIR-Arxiv	MIMIR-PileCC	prior	0.499 / 1.000	$0.248^{\dagger}$ / 1.000
Pythia-2.8B	MIMIR-PileCC	MIMIR-Arxiv	all	0.499 / 1.000	**0.011** / 1.000
Pythia-2.8B	MIMIR-PileCC	MIMIR-Arxiv	prior	0.499 / 1.000	**0.015** / 1.000

Table 2 reports both successful and failed external-null calibration cases. On GPT-2, AKTD (E) sharply reduces empirical FDP relative to ordinary KTD. The prior-0.75 null-count bound is less conservative than the all-null bound and recovers more power while staying below the target. Across Pythia-410M/1B/2.8B, the successful direction is consistently MIMIR-PileCC calibrated by MIMIR-Arxiv: AKTD (E) has empirical FDP between 0.006 and 0.016 with essentially full power. However, MIMIR-Arxiv calibrated by MIMIR-PileCC, consistently exceeds the target, with empirical FDP from 0.182 to 0.312. Thus the larger Pythia runs support the null-mismatch interpretation rather than revealing an isolated irregularity.

Limitations and scope. The failed Pythia/MIMIR-Arxiv rows show that the matched-null requirement for AKTD (E) is substantive. Figure 1 visualizes the penalty-dominance behavior for GPT-2

Table 3: External-null calibrated score-only baselines versus AKTD (E) on MIMIR at $q = 0.1$. Score-only entries report the best result among LOSS, MinK%, and Zlib. The n_0 bound is either all-null or a prior bound $n_0 \leq 0.75n$. Bold marks the highest-power method among target-satisfying methods in each row; † marks FDP $> q$.

Candidate set	Model	n_0 bound	Best external score	Score Emp. FDP↓ / Power↑	AKTD (E) Emp. FDP↓ / Power↑
MIMIR-Arxiv	GPT-2	all	MinK% (10%)	**0.037 / 0.994**	0.042 / 0.458
MIMIR-Arxiv	GPT-2	prior	MinK% (10%)	**0.048 / 0.998**	0.055 / 0.620
MIMIR-Arxiv	Pythia-410M	all	MinK% (5%)	**0.092 / 1.000**	0.206† / 1.000
MIMIR-Arxiv	Pythia-410M	prior	MinK% (10%)	0.116† / 1.000	0.295† / 1.000
MIMIR-PileCC	GPT-2	all	MinK% (20%)	**0.082 / 0.955**	0.048 / 0.397
MIMIR-PileCC	GPT-2	prior	MinK% (20%)	**0.088 / 0.959**	0.065 / 0.458
MIMIR-PileCC	Pythia-410M	all	Zlib (ppl_zlib)	**0.001 / 1.000**	0.010 / 0.999
MIMIR-PileCC	Pythia-410M	prior	Zlib (ppl_zlib)	**0.001 / 1.000**	0.016 / 0.999

and Pythia-410M, and Appendix C.4 extends the same diagnostic to Pythia-1B and Pythia-2.8B. These diagnostics show that, when the external pool underestimates the evaluated null asymmetry, AKTD (E) can choose an overly liberal threshold and inflate FDP.

Appendices C.5 and C.6 clarify two practical boundary cases of AKTD (E). Shifted WikiMIA external null pools make it over-conservative with no discoveries at $q = 0.1$, while certified penalties are valid but currently too conservative for useful power.

5.4 Comparison with Score Baselines

We finally compare AKTD with score-based baselines. This comparison is not intended as a pure ranking-score leaderboard. LOSS, MinK%, and Zlib directly rank samples by membership evidence, whereas AKTD addresses whether knockoff thresholding remains valid when generated knockoffs are only approximately exchangeable.

For a fair contextual comparison, we also calibrate each score baseline using an external null pool. Intuitively, the threshold is chosen so that the estimated number of external-null samples exceeding the threshold is small relative to the number of selected candidates. This calibration step does not use evaluated candidate-set labels.

The best score entry in Table 3 is selected post hoc among LOSS, MinK%, and Zlib for each setting, so it should be read as an optimistic diagnostic reference rather than a fully label-free model selection.

The results are deliberately favorable to score baselines. The displayed baseline is selected post hoc and can be a very strong ranker. This supports rather than replaces the role of AKTD. Matched external nulls can help calibrate any strong statistic, while AKTD targets the distinct thresholding failure of ordinary KTD under null-tail asymmetry. In the Pythia MIMIR-Arxiv rows, its failure is consistent with the external-null mismatch diagnostics, reinforcing that reliable external-null transfer requires matched null statistics.

6 Conclusion

Approximate exchangeability creates a concrete failure mode for knockoff-based training data detection: violations of exact null sign symmetry induce null-tail asymmetry, causing the standard knockoff+ threshold to become anti-conservative. A pathwise FDP bound quantifies this deviation and explains the empirical FDP inflation observed for ordinary KTD. Therefore, we proposed Asymmetry-adjusted KTD (AKTD), which corrects the threshold using an estimated asymmetry penalty. Oracle and cross-fit variants validate the mechanism, while external-null AKTD provides a candidate-set label-free calibration route when a sufficiently matched null-only calibration pool is available. Experiments on WikiMIA and MIMIR show that AKTD restores controlled selection in diagnostic settings and succeeds in several external-null MIMIR settings. More broadly, AKTD provides the pathwise diagnosis and correction framework for knockoff-based LLM training data detection under imperfect text exchangeability.

Funding. This work was supported in part by the National Natural Science Foundation of China under Grants 62502517, 62271485, 62303462, and 62306338; Development Plan for Youth Innovation Teams in Higher Education Institutions of Shandong Province No. 2024KJH076.

Competing interests. The authors declare no competing interests.

References

Rina Foygel Barber and Emmanuel J Candes. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

Rina Foygel Barber, Emmanuel J Candès, and Richard J Samworth. Robust inference with knockoffs. *The Annals of Statistics*, 48(3):1409–1431, 2020.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE, 2022a.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022b.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, memory: An archaeology of books known to chatgpt/gpt-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, 2023.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*, 2024.

Yingying Fan, Lan Gao, and Jinchi Lv. Ark: Robust knockoffs inference with coupling. *The Annals of Statistics*, 53(2):749–773, 2025a.

Yingying Fan, Lan Gao, Jinchi Lv, and Xiaocong Xu. Asymptotic fdr control with model-x knockoffs: Is moments matching sufficient? *arXiv preprint arXiv:2502.05969*, 2025b.

Zirui Hu, Yingjie Wang, Zheng Zhang, Hong Chen, and Dacheng Tao. A statistical approach for controlled training data detection. In *The Thirteenth International Conference on Learning Representations*, 2025.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR, 2022.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092, 2024.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, 2023.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. Recall: Membership inference via relative conditional log-likelihoods. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8671–8689, 2024.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

APPENDIX

A Proofs for AKTD

This appendix provides the theoretical details for Section 4. We use the notation from the main text. For $t > 0$, recall

$$R(t) = \sum_{j=1}^{n} \mathbf{1}\{W_j \geq t\}, \qquad N(t) = \sum_{j=1}^{n} \mathbf{1}\{W_j \leq -t\},$$

and the null tail counts

$$V_+(t) = \sum_{j \in \mathcal{H}_0} \mathbf{1}\{W_j \geq t\}, \qquad V_-(t) = \sum_{j \in \mathcal{H}_0} \mathbf{1}\{W_j \leq -t\}.$$

We also write

$$D_0(t) = \{V_+(t) - V_-(t)\}_+, \qquad \Delta_0(t) = \frac{D_0(t)}{R(t) \vee 1}.$$

For adjusted procedures, we write q_{int} for the internal thresholding level. The symbol q is reserved for the final target FDR level when level splitting is used.

A.1 Pairwise Exchangeability and Null Sign Symmetry

We first record a sufficient condition under which text-level exchangeability implies sign symmetry of the induced statistic. This result connects the notation to the classical knockoff intuition. It does not claim that arbitrary multi-knockoff aggregations automatically satisfy sign symmetry.

Lemma 1 (Pairwise exchangeability with antisymmetric statistics) *Fix a null index $j \in \mathcal{H}_0$. Suppose that, conditional on all remaining samples, knockoffs, and model information used by the scoring rule,*

$$(X_j, \widetilde{X}_j) \overset{d}{=} (\widetilde{X}_j, X_j).$$

Suppose also that the statistic is antisymmetric under swapping the original and knockoff sample:

$$W_j(X_j, \widetilde{X}_j) = -W_j(\widetilde{X}_j, X_j).$$

Then

$$W_j \overset{d}{=} -W_j$$

under the null. In particular, for every $t > 0$,

$$\mathbb{P}(W_j \geq t) = \mathbb{P}(W_j \leq -t),$$

up to the convention for atoms at zero.

Proof 1 *By conditional exchangeability,*

$$W_j(X_j, \widetilde{X}_j) \overset{d}{=} W_j(\widetilde{X}_j, X_j).$$

By antisymmetry,

$$W_j(\widetilde{X}_j, X_j) = -W_j(X_j, \widetilde{X}_j).$$

Combining the two displays gives $W_j \overset{d}{=} -W_j$. The equality of the positive and negative tails follows immediately.

If the joint null law is invariant under arbitrary swaps of any subset of null original–knockoff pairs, and if the statistic is antisymmetric for each such pair, then the same argument applies jointly. Conditional on the null magnitudes and the non-null statistics, every null sign pattern has the same probability as its sign-flipped counterpart, yielding the statistic-level null sign-symmetry condition used by the ordinary knockoff+ proof.

For the multi-knockoff statistic used in this paper,

$$W_j = s(X_j; \theta) - \frac{1}{m} \sum_{\ell=1}^{m} s(\widetilde{X}_j^{(\ell)}; \theta),$$

pairwise antisymmetry under a single swap need not hold when $m > 1$. We therefore state the main analysis directly at the level of the induced statistic vector $W = (W_1, \ldots, W_n)$ through null sign symmetry and realized null-tail asymmetry.

A.2 Negative-Tail Surrogate under Null Sign Symmetry

The following is the standard knockoff+ sign-flip lemma in our notation.

Lemma 2 (Knockoff+ sign-flip lemma) *Assume null sign symmetry: conditional on the magnitudes $\{|W_j| : j \in [n]\}$ and the non-null statistics $\{W_j : j \in \mathcal{H}_1\}$, the null signs are independent Rademacher random variables. Let*

$$\tau = \inf\left\{t \in \mathcal{T} : \frac{1 + N(t)}{R(t) \vee 1} \leq q\right\}$$

be the ordinary knockoff+ threshold. Then

$$\mathbb{E}\left[\frac{V_+(\tau)}{1 + V_-(\tau)}\right] \leq 1.$$

Proof 2 *Conditional on the magnitudes and non-null statistics, order the null hypotheses by decreasing magnitude:*

$$A_{(1)} \geq A_{(2)} \geq \cdots \geq A_{(n_0)}, \qquad A_j = |W_j|, \qquad n_0 = |\mathcal{H}_0|.$$

For any threshold t, the null hypotheses with $|W_j| \geq t$ form an initial segment of this ordering. Under null sign symmetry, the signs along this ordered sequence are independent fair coin flips. The knockoff+ threshold is a stopping rule with respect to the filtration generated by revealing these signs together with the fixed magnitudes and non-null statistics. The standard knockoff+ optional-stopping argument for fair signs gives

$$\mathbb{E}\left[\frac{\#\{\textit{positive null signs before the stopping threshold}\}}{1 + \#\{\textit{negative null signs before the stopping threshold}\}}\right] \leq 1.$$

The numerator and denominator are $V_+(\tau)$ and $1 + V_-(\tau)$, respectively.

A.3 Ordinary KTD under Exact Null Sign Symmetry

Proposition 3 (Ordinary KTD FDR control) *Under the null sign-symmetry condition in Lemma 2, the ordinary KTD selected set*

$$\widehat{S}_{\mathrm{KTD}} = \{j : W_j \geq \tau_{\mathrm{KTD}}\}$$

satisfies

$$\mathrm{FDR}_{\mathrm{KTD}} \leq q.$$

Proof 3 *If $\tau_{\mathrm{KTD}} = \infty$, then $\widehat{S}_{\mathrm{KTD}} = \emptyset$ and the FDP is zero. Otherwise,*

$$\mathrm{FDP}_{\mathrm{KTD}} = \frac{|\widehat{S}_{\mathrm{KTD}} \cap \mathcal{H}_0|}{|\widehat{S}_{\mathrm{KTD}}| \vee 1} = \frac{V_+(\tau_{\mathrm{KTD}})}{R(\tau_{\mathrm{KTD}}) \vee 1}.$$

By definition of τ_{KTD},

$$\frac{1 + N(\tau_{\mathrm{KTD}})}{R(\tau_{\mathrm{KTD}}) \vee 1} \leq q.$$

Since $V_-(t) \leq N(t)$ for every t,

$$\frac{1 + V_-(\tau_{\mathrm{KTD}})}{R(\tau_{\mathrm{KTD}}) \vee 1} \leq \frac{1 + N(\tau_{\mathrm{KTD}})}{R(\tau_{\mathrm{KTD}}) \vee 1} \leq q.$$

Therefore,

$$\mathrm{FDP}_{\mathrm{KTD}} = \frac{V_+(\tau_{\mathrm{KTD}})}{1 + V_-(\tau_{\mathrm{KTD}})} \cdot \frac{1 + V_-(\tau_{\mathrm{KTD}})}{R(\tau_{\mathrm{KTD}}) \vee 1} \leq q\, \frac{V_+(\tau_{\mathrm{KTD}})}{1 + V_-(\tau_{\mathrm{KTD}})}.$$

Taking expectations and applying Lemma 2 gives

$$\mathrm{FDR}_{\mathrm{KTD}} = \mathbb{E}[\mathrm{FDP}_{\mathrm{KTD}}] \leq q\, \mathbb{E}\left[\frac{V_+(\tau_{\mathrm{KTD}})}{1 + V_-(\tau_{\mathrm{KTD}})}\right] \leq q.$$

A.4 Proof of the Pathwise FDP Deviation Bound

We prove Proposition 1. The result is deterministic and does not require sign symmetry.

Proof 4 *If $\tau_{\mathrm{KTD}} = \infty$, then $\widehat{S}_{\mathrm{KTD}} = \emptyset$ and the FDP is zero. Assume $\tau_{\mathrm{KTD}} < \infty$. For any threshold t,*

$$V_+(t) \le V_-(t) + \{V_+(t) - V_-(t)\}_+ = V_-(t) + D_0(t).$$

Dividing by $R(t) \vee 1$ gives

$$\frac{V_+(t)}{R(t) \vee 1} \le \frac{V_-(t)}{R(t) \vee 1} + \frac{D_0(t)}{R(t) \vee 1} = \frac{V_-(t)}{R(t) \vee 1} + \Delta_0(t).$$

Evaluating at $t = \tau_{\mathrm{KTD}}$,

$$\mathrm{FDP}_{\mathrm{KTD}}(W) = \frac{V_+(\tau_{\mathrm{KTD}})}{R(\tau_{\mathrm{KTD}}) \vee 1} \le \frac{V_-(\tau_{\mathrm{KTD}})}{R(\tau_{\mathrm{KTD}}) \vee 1} + \Delta_0(\tau_{\mathrm{KTD}}).$$

Since $V_-(t) \le N(t)$ for all t,

$$\frac{V_-(\tau_{\mathrm{KTD}})}{R(\tau_{\mathrm{KTD}}) \vee 1} \le \frac{N(\tau_{\mathrm{KTD}})}{R(\tau_{\mathrm{KTD}}) \vee 1} \le \frac{1 + N(\tau_{\mathrm{KTD}})}{R(\tau_{\mathrm{KTD}}) \vee 1} \le q.$$

The last inequality follows from the definition of τ_{KTD}. Thus

$$\mathrm{FDP}_{\mathrm{KTD}}(W) \le q + \Delta_0(\tau_{\mathrm{KTD}}) \le q + \Delta_0^\star.$$

Taking expectations gives

$$\mathrm{FDR}_{\mathrm{KTD}} \le q + \mathbb{E}[\Delta_0^\star].$$

A.5 Proof of the AKTD Adjusted-Threshold Certificate

We prove Proposition 2, where the adjusted procedure is run at internal level q_{int}.

Proof 5 *If $\tau_B = \infty$, then $\widehat{S}_B = \emptyset$ and the FDP is zero. Assume $\tau_B < \infty$. By assumption,*

$$B(t) \ge D_0(t) = \{V_+(t) - V_-(t)\}_+ \qquad \textit{for all } t \in \mathcal{T}.$$

Therefore, for every $t \in \mathcal{T}$,

$$V_+(t) \le V_-(t) + D_0(t) \le V_-(t) + B(t) \le N(t) + B(t).$$

Evaluating at $t = \tau_B$,

$$\mathrm{FDP}_B(W) = \frac{V_+(\tau_B)}{R(\tau_B) \vee 1} \le \frac{N(\tau_B) + B(\tau_B)}{R(\tau_B) \vee 1} \le \frac{1 + N(\tau_B) + B(\tau_B)}{R(\tau_B) \vee 1}.$$

By the definition of τ_B,

$$\frac{1 + N(\tau_B) + B(\tau_B)}{R(\tau_B) \vee 1} \le q_{\mathrm{int}}.$$

Thus $\mathrm{FDP}_B(W) \le q_{\mathrm{int}}$.

A.6 Uniform Concentration for External-Null Calibration

We now establish the concentration statement used in the external-null certificate. Let F_0 denote the common null distribution of the statistic under the same-null calibration assumption. Let $C_1, \dots, C_M \overset{\text{i.i.d.}}{\sim} F_0$ be the external null calibration statistics, independent of the evaluated null statistics $\{W_j : j \in \mathcal{H}_0\}$, which are also assumed to be draws from F_0. Define

$$d_0(t) = \mathbb{P}_{C \sim F_0}(C \ge t) - \mathbb{P}_{C \sim F_0}(C \le -t), \qquad a_0(t) = \{d_0(t)\}_+,$$

and

$$\widehat{d}_M(t) = \frac{1}{M} \sum_{i=1}^{M} \left[\mathbf{1}\{C_i \ge t\} - \mathbf{1}\{C_i \le -t\}\right], \qquad \widehat{a}_M(t) = \{\widehat{d}_M(t)\}_+.$$

Lemma 3 (Uniform external-null concentration) *Let $\alpha \in (0,1)$ and define*

$$\eta_M(\alpha) = \sqrt{\frac{2\log(4/\alpha)}{M}}, \qquad \eta_n(\alpha) = \sqrt{2\bar{n}_0 \log(4/\alpha)}.$$

Assume $|\mathcal{H}_0| = n_0 \le \bar{n}_0$. With probability at least $1-\alpha$, simultaneously for all $t \in \mathbb{R}$,

$$D_0(t) \le \bar{n}_0\{\widehat{a}_M(t) + \eta_M(\alpha)\} + \eta_n(\alpha).$$

Proof 6 *If $n_0 = 0$, then $D_0(t) = 0$ for all t and the claim is immediate. Assume $n_0 \ge 1$.*

Let $\widehat{F}_M$ be the empirical distribution function of $C_1, \dots, C_M$. By the Dvoretzky–Kiefer–Wolfowitz inequality, with probability at least $1-\alpha/2$,

$$\sup_x |\widehat{F}_M(x) - F_0(x)| \le \frac{1}{2}\eta_M(\alpha).$$

On this event,

$$|\widehat{d}_M(t) - d_0(t)| \le 2\sup_x |\widehat{F}_M(x) - F_0(x)| \le \eta_M(\alpha) \qquad \textit{for all } t.$$

Therefore,

$$a_0(t) = \{d_0(t)\}_+ \le \{\widehat{d}_M(t) + \eta_M(\alpha)\}_+ \le \widehat{a}_M(t) + \eta_M(\alpha) \qquad \textit{for all } t.$$

Let $\widehat{F}_{0,n}$ be the empirical distribution function of the evaluated null statistics $\{W_j : j \in \mathcal{H}_0\}$. Applying the same DKW inequality, with probability at least $1-\alpha/2$,

$$\sup_x |\widehat{F}_{0,n}(x) - F_0(x)| \le \frac{1}{2}\sqrt{\frac{2\log(4/\alpha)}{n_0}}.$$

Thus, uniformly over t,

$$\left|\frac{V_+(t) - V_-(t)}{n_0} - d_0(t)\right| \le \sqrt{\frac{2\log(4/\alpha)}{n_0}}.$$

Multiplying by n_0 gives

$$V_+(t) - V_-(t) \le n_0 d_0(t) + \sqrt{2n_0\log(4/\alpha)} \le n_0 d_0(t) + \eta_n(\alpha).$$

Taking positive parts and using $\eta_n(\alpha) \ge 0$,

$$D_0(t) = \{V_+(t) - V_-(t)\}_+ \le n_0 a_0(t) + \eta_n(\alpha).$$

Combining this with the calibration bound on $a_0(t)$ and $n_0 \le \bar{n}_0$ yields

$$D_0(t) \le \bar{n}_0\{\widehat{a}_M(t) + \eta_M(\alpha)\} + \eta_n(\alpha)$$

for all t. A union bound over the two DKW events completes the proof.

The concentration statement is uniform over all real thresholds, not merely a fixed grid. It therefore applies to the data-dependent threshold selected by Algorithm 1.

A.7 Proof of the External-Null AKTD Certificate

We prove Theorem 1.

Proof 7 *Define*

$$B_\alpha(t) = \bar{n}_0\{\widehat{a}_M(t) + \eta_M(\alpha)\} + \eta_n(\alpha).$$

By Lemma 3, with probability at least $1-\alpha$,

$$B_\alpha(t) \ge D_0(t) \qquad \textit{for all } t \in \mathbb{R}.$$

In particular, this domination holds for every $t \in \mathcal{T}$ and at the data-dependent threshold selected by Algorithm 1. On this event, Proposition 2 applies with $B(t) = B_\alpha(t)$, so

$$\mathrm{FDP}_{\mathrm{ext}}(W) \le q_{\mathrm{int}}.$$

Therefore,

$$\mathbb{P}\left(\mathrm{FDP}_{\mathrm{ext}}(W) \le q_{\mathrm{int}}\right) \ge 1-\alpha.$$

Since $0 \le \mathrm{FDP}_{\mathrm{ext}} \le 1$,

$$\mathrm{FDR}_{\mathrm{ext}} \le q_{\mathrm{int}}(1-\alpha) + \alpha \le q_{\mathrm{int}} + \alpha.$$

This proves the theorem.

A.8 Exact FDR Control from a High-Probability Certificate

The external-null theorem gives a high-probability FDP certificate at the internal level q_{int} and hence

$$\mathrm{FDR}_{\mathrm{ext}} \le q_{\mathrm{int}}(1-\alpha) + \alpha.$$

To control FDR at a target level q, one can choose a smaller internal level.

Corollary 1 (External-null FDR control by level splitting) *Suppose the external-null AKTD certificate holds with probability at least $1-\alpha$ at internal level q_{int}. If*

$$q_{\mathrm{int}} = \frac{q-\alpha}{1-\alpha},$$

where $0 < \alpha < q < 1$, then

$$\mathrm{FDR}_{\mathrm{ext}} \le q.$$

The simpler choice $q_{\mathrm{int}} = q - \alpha$ also gives $\mathrm{FDR}_{\mathrm{ext}} \le q$ whenever $0 < \alpha \le q$.

Proof 8 *Running the procedure at internal level q_0 gives*

$$\mathbb{P}(\mathrm{FDP}_{\mathrm{ext}} \le q_0) \ge 1-\alpha.$$

Since the FDP is bounded by one,

$$\mathrm{FDR}_{\mathrm{ext}} \le q_0(1-\alpha) + \alpha.$$

For $q_0 = (q-\alpha)/(1-\alpha)$, the right-hand side equals q.

For the simpler choice $q_0 = q - \alpha$,

$$(q-\alpha)(1-\alpha) + \alpha = q - q\alpha + \alpha^2 \le q$$

whenever $\alpha \le q$.

A.9 Oracle, Cross-Fit, and Empirical Penalties

Oracle, cross-fit, and bootstrap AKTD differ only in the construction of the penalty $B(t)$ used by Algorithm 1. The oracle penalty is

$$B_{\mathrm{oracle}}(t) = D_0(t) = \{V_+(t) - V_-(t)\}_+.$$

It satisfies the domination condition with equality, so Proposition 2 gives pathwise FDP control:

$$\mathrm{FDP}_{\mathrm{oracle}}(W) \le q_{\mathrm{int}}.$$

This variant requires true null labels and is used only as a mechanism check.

Cross-fit AKTD estimates $D_0(t)$ from held-out benchmark null folds. If the resulting estimate $\widehat{B}_{\mathrm{cf}}(t)$ satisfies

$$\widehat{B}_{\mathrm{cf}}(t) \ge D_0(t) \qquad \text{for all } t \in \mathcal{T}$$

on a given realization, then Proposition 2 implies $\mathrm{FDP}_{\mathrm{cf}}(W) \le q_{\mathrm{int}}$ on that realization. In the main experiments, cross-fit AKTD is used as a label-assisted diagnostic protocol rather than as a deployable guarantee.

The certified external-null penalty $B_\alpha(t)$ provides a finite-sample certificate under the same-null assumption. Bootstrap or quantile penalties are empirical estimates of the same count-scale asymmetry. They can improve power, but their guarantee is conditional on whether they dominate $D_0(t)$ along the realized threshold path. If the domination condition fails, they should be read as empirical calibration procedures rather than finite-sample certified procedures.

B Additional Algorithms

The main text presents the generic AKTD thresholding step (Algorithm 1), since all AKTD variants share the same selection rule once a penalty function $B(t)$ has been specified. Algorithms 2 and 3 provide implementation-level details for two representative cases: ordinary KTD and certified external-null AKTD.

Algorithm 2 Ordinary KTD

Require: Candidate texts $\{X_j\}_{j=1}^n$, model f_θ, score s, number of knockoffs m, target level q

1: Generate knockoffs $\{\widetilde{X}_j^{(\ell)}\}_{\ell=1}^m$ for each candidate text X_j
2: Compute the multi-knockoff statistic

$$W_j = s(X_j;\theta) - \frac{1}{m}\sum_{\ell=1}^{m} s(\widetilde{X}_j^{(\ell)};\theta), \qquad j \in [n].$$

3: Form the threshold grid

$$\mathcal{T} = \{|W_j| : |W_j| > 0,\ j \in [n]\}.$$

4: If $\mathcal{T} = \emptyset$, return $\emptyset$
5: For all $t \in \mathcal{T}$, compute

$$R(t) = \sum_{j=1}^{n} \mathbf{1}\{W_j \geq t\}, \qquad N(t) = \sum_{j=1}^{n} \mathbf{1}\{W_j \leq -t\}.$$

6: Compute the ordinary knockoff+ threshold

$$\tau_{\mathrm{KTD}} = \inf\left\{t \in \mathcal{T} : \frac{1+N(t)}{R(t)\vee 1} \leq q\right\}.$$

7: If no such t exists, set $\tau_{\mathrm{KTD}} = \infty$
8: **return**

$$\widehat{S}_{\mathrm{KTD}} = \{j : W_j \geq \tau_{\mathrm{KTD}}\}.$$

Ordinary KTD first constructs the statistic vector W from original and knockoff texts, and then applies the knockoff+ threshold. It can be viewed as the unadjusted special case of Algorithm 1 with $B(t) \equiv 0$.

Certified external-null AKTD assumes that the candidate statistics W_j and external null statistics C_i have already been computed using the same score and knockoff-generation pipeline. It constructs the finite-sample penalty $B_\alpha(t)$ from the external null pool and then calls the generic AKTD thresholding rule at internal level q_{int}. The procedure does not use labels from the evaluated candidate set. Its

Algorithm 3 Certified External-null AKTD

Require: Candidate statistics $\{W_j\}_{j=1}^n$, external null statistics $\{C_i\}_{i=1}^M$, internal level q_{int}, confidence level α, null-count upper bound $\bar{n}_0$

1: Form the threshold grid

$$\mathcal{T} = \{|W_j| : |W_j| > 0,\ j \in [n]\}.$$

2: If $\mathcal{T} = \emptyset$, return $\emptyset$
3: For every $t \in \mathcal{T}$, compute the external null-tail asymmetry estimate

$$\widehat{a}_M(t) = \left\{\frac{1}{M}\sum_{i=1}^{M}[\mathbf{1}\{C_i \geq t\} - \mathbf{1}\{C_i \leq -t\}]\right\}_+.$$

4: Set

$$\eta_M(\alpha) = \sqrt{\frac{2\log(4/\alpha)}{M}}, \qquad \eta_n(\alpha) = \sqrt{2\bar{n}_0\log(4/\alpha)}.$$

5: Construct the certified penalty

$$B_\alpha(t) = \bar{n}_0\{\widehat{a}_M(t) + \eta_M(\alpha)\} + \eta_n(\alpha), \qquad t \in \mathcal{T}.$$

6: Run Algorithm 1 with internal level q_{int} and penalty $B(t) = B_\alpha(t)$
7: **return** the selected set $\widehat{S}_{\mathrm{ext}}$

Table 4: Map of appendix experiments and their role in supporting the main paper.

Appendix section	Purpose
Appendix C.1	Tests the pathwise diagnosis across several target levels on WikiMIA/GPT-2.
Appendix C.2	Compares raw, smoothed, and bootstrap asymmetry estimators.
Appendix C.3	Compares gradient norm and log-probability statistics on MIMIR-Arxiv/GPT-2.
Appendix C.4	Checks external-null penalty dominance for the Pythia Models.
Appendix C.5	Compares the theorem-certified external-null penalty with the bootstrap penalty.
Appendix C.6	Tests available external null pools for WikiMIA/GPT-2.

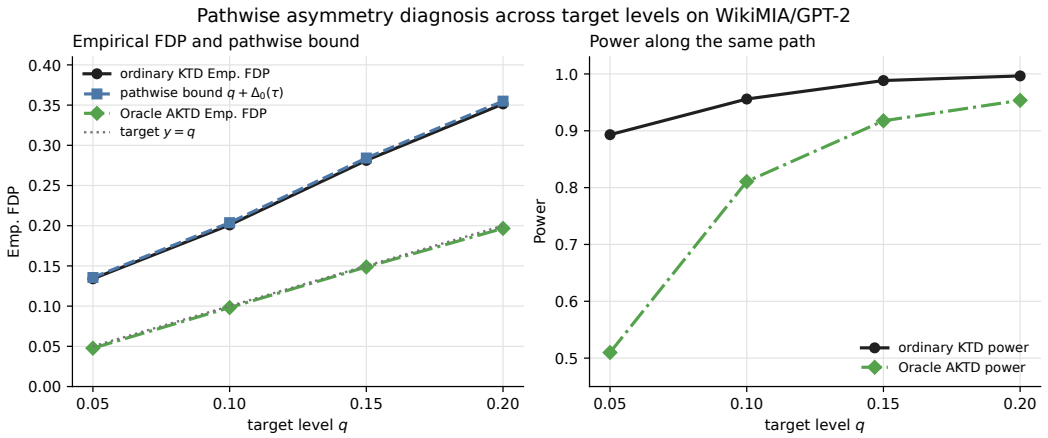

Figure 2: Pathwise asymmetry diagnosis across target levels on WikiMIA/GPT-2. Left: the pathwise bound $q + \Delta_0(\tau)$ tracks the empirical FDP of ordinary KTD across target levels, while Oracle AKTD restores empirical FDP to the target line $y = q$. Right: Oracle AKTD achieves this correction at the cost of reduced power.

certificate requires a valid upper bound $\bar{n}_0$ and the same-null condition in Theorem 1: the external null statistics must be representative of the evaluated null statistics.

Bootstrap and quantile penalties used in the empirical experiments follow the same implementation pattern. They replace $B_\alpha(t)$ with an empirical penalty and then run Algorithm 1. These variants are not finite-sample certified unless their penalties dominate the realized asymmetry as required by Proposition 2.

Oracle and cross-fit AKTD do not require separate pseudocode. Oracle AKTD uses Algorithm 1 with

$$B(t) = D_0(t) = \{V_+(t) - V_-(t)\}_+,$$

and is used only for mechanism validation because it requires true null labels. Cross-fit AKTD estimates $B(t)$ from held-out benchmark null folds and then applies the same generic thresholding rule. It is a label-assisted diagnostic protocol rather than a deployable auditing procedure.

C Additional Experimental Details

This appendix provides additional evidence for the experimental claims in Section 5. Table 4 summarizes the appendix experiments and clarifies how they support the main text. The appendix focuses on five issues: whether the pathwise diagnosis holds beyond the single $q = 0.1$ operating point, how practical estimator and statistic choices affect AKTD, whether the Pythia external-null pattern persists at larger checkpoints, how the theorem-certified penalty compares with the bootstrap penalty, and how external-null calibration behaves for AKTD and score-only baselines.

Unless stated otherwise, all false-discovery quantities in this appendix are empirical FDP values computed on labeled benchmark candidate sets. They are realized benchmark diagnostics, not direct estimates of expectation-level FDR.

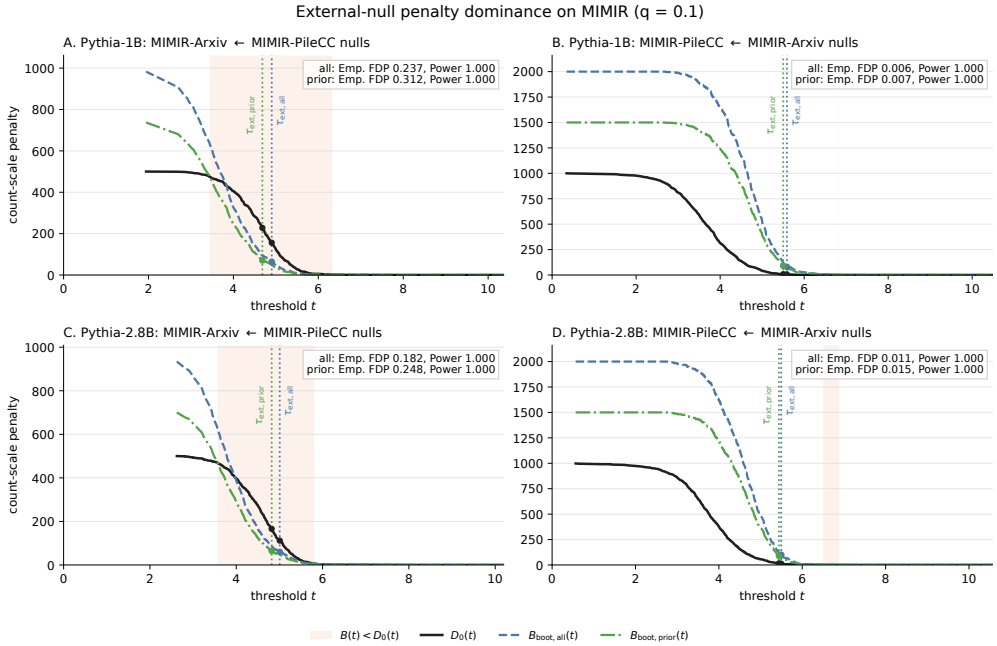

Figure 3: Penalty dominance diagnostics for fine-tuned Pythia-1B and Pythia-2.8B at $q = 0.1$. The plotting convention matches Figure 1: black curves show the realized asymmetry $D_0(t)$, blue and green curves show the external-null bootstrap penalties, dotted lines mark selected thresholds, and shaded regions indicate where an external penalty falls below $D_0(t)$.

Table 5: Asymmetry estimator ablation on WikiMIA/GPT-2 at $q = 0.1$. Lower empirical FDP is better, while higher power is preferred.

Estimator	Fixed 25% calibration Mean Emp. FDP↓ / Power↑	Fixed 25% calibration Max Emp. FDP↓	5-fold cross-fit Emp. FDP↓ / Power↑
Raw	0.102 / 0.877	0.140	0.101 / 0.792
Smooth q=0.8	**0.094 / 0.862**	0.136	0.092 / 0.720
Smooth q=0.9	0.093 / 0.860	0.134	0.091 / 0.686
Bootstrap q=0.8	0.089 / 0.846	0.125	**0.092 / 0.734**
Bootstrap q=0.9	0.086 / 0.819	0.116	0.082 / 0.689

C.1 Pathwise Diagnosis across Target Levels

The main text reports the pathwise asymmetry diagnosis at $q = 0.1$. Here we check whether the agreement between empirical FDP and the pathwise bound holds across several target levels. Figure 2 shows the WikiMIA/GPT-2 q-grid for $q \in \{0.05, 0.10, 0.15, 0.20\}$.

Across all four target levels, the pathwise bound remains close to the empirical FDP of ordinary KTD. Oracle AKTD tracks the target line, confirming that realized null-tail asymmetry explains the observed FDP inflation.

C.2 Practical Asymmetry Estimator Ablation

The oracle penalty validates the mechanism but uses true null labels. We therefore compare practical estimators of the null-tail asymmetry on WikiMIA/GPT-2 in Table 5. All variants use the gradient-norm statistic and ten knockoffs per sample. Fixed calibration holds out 25% of benchmark nulls and averages over five seeds. Cross-fit uses five folds and pools discoveries.

Raw calibration gives high power but stays close to the target boundary and has larger seed-level deviations. Bootstrap q=0.8 gives the best cross-fit control–power tradeoff among these settings, and is therefore used as the default practical estimator in the main experiments.

Table 6: Statistic ablation on MIMIR-Arxiv/GPT-2 at $q = 0.1$. Lower empirical FDP and lower $\Delta_0^\star$ are better, while higher AUC and power are preferred.

Statistic	AUC↑	$\Delta_0^\star$ ↓	KTD Emp. FDP↓ / Power↑	AKTD (O) Emp. FDP↓ / Power↑	AKTD (C) Emp. FDP↓ / Power↑
Gradient norm	0.813	1.000	0.499 / 1.000	0.182† / 0.360	0.000 / 0.000
Log-probability	**0.917**	**0.500**	0.499 / 1.000	**0.095 / 0.724**	**0.073 / 0.684**

C.3 Statistic Ablation on MIMIR-Arxiv/GPT-2

This ablation tests whether AKTD depends on the underlying KTD statistic. Table 6 compares gradient norm and log-probability on MIMIR-Arxiv/GPT-2 using the same bootstrap q=0.8 asymmetry estimator.

Ordinary KTD has severe empirical FDP inflation for both statistics. However, the adjusted procedure retains useful power only with the log-probability statistic. The gradient statistic is too unstable under the MIMIR paraphrase protocol: it induces larger asymmetry and yields no cross-fit discoveries at $q = 0.1$. The log-probability statistic has higher AUC, smaller realized asymmetry, and useful adjusted power, so we use it for the MIMIR main experiments.

C.4 Pythia Scale Diagnostics for External-Null Calibration

Figure 3 repeats the external-null penalty diagnostic for fine-tuned Pythia-1B and Pythia-2.8B. Both model sizes show the same directional pattern. The MIMIR-PileCC candidate set is well controlled using MIMIR-Arxiv nulls, whereas the MIMIR-Arxiv candidate set remains anti-conservative when calibrated using MIMIR-PileCC nulls. This supports the interpretation that the failure is driven by external-null mismatch rather than insufficient target-model scale.

C.5 Certified versus Bootstrap External-Null Penalties

The finite-sample theory provides a certified external-null penalty via uniform concentration, whereas the main experiments use a bootstrap-quantile penalty for power. Table 7 compares the two penalties on the MIMIR external-null settings. Certified rows use $\alpha = 0.05$. Bootstrap rows use 300 bootstrap repetitions, quantile 0.8, and smoothing window 1.

Table 7: Certified versus bootstrap external-null penalties at $q = 0.1$. The certified column reports the minimum certified threshold ratio; all values remain above q, so the certified rule makes no discoveries. The n_0 bound is either all-null or a prior bound $n_0 \leq 0.75n$. Bootstrap rows report empirical FDP/power and selected counts; † marks FDP $> q$.

Model	Candidate ← external null	n_0 bound	Certified min ratio↓	Bootstrap Emp. FDP↓ / Power↑	Bootstrap selected↑
GPT-2	MIMIR-Arxiv ← MIMIR-PileCC	all	0.762	**0.042 / 0.458**	239
GPT-2	MIMIR-Arxiv ← MIMIR-PileCC	prior	0.602	**0.055 / 0.620**	328
GPT-2	MIMIR-PileCC ← MIMIR-Arxiv	all	1.211	**0.048 / 0.397**	417
GPT-2	MIMIR-PileCC ← MIMIR-Arxiv	prior	0.940	**0.065 / 0.458**	490
Pythia-410M	MIMIR-Arxiv ← MIMIR-PileCC	all	0.541	0.206† / 1.000	630
Pythia-410M	MIMIR-Arxiv ← MIMIR-PileCC	prior	0.429	0.295† / 1.000	709
Pythia-410M	MIMIR-PileCC ← MIMIR-Arxiv	all	0.660	**0.010 / 0.999**	1009
Pythia-410M	MIMIR-PileCC ← MIMIR-Arxiv	prior	0.521	**0.016 / 0.999**	1015
Pythia-1B	MIMIR-Arxiv ← MIMIR-PileCC	all	0.287	0.237† / 1.000	655
Pythia-1B	MIMIR-Arxiv ← MIMIR-PileCC	prior	0.216	0.312† / 1.000	727
Pythia-1B	MIMIR-PileCC ← MIMIR-Arxiv	all	0.426	**0.006 / 1.000**	1006
Pythia-1B	MIMIR-PileCC ← MIMIR-Arxiv	prior	0.320	**0.007 / 1.000**	1007
Pythia-2.8B	MIMIR-Arxiv ← MIMIR-PileCC	all	0.293	0.182† / 1.000	611
Pythia-2.8B	MIMIR-Arxiv ← MIMIR-PileCC	prior	0.220	0.248† / 1.000	665
Pythia-2.8B	MIMIR-PileCC ← MIMIR-Arxiv	all	0.426	**0.011 / 1.000**	1011
Pythia-2.8B	MIMIR-PileCC ← MIMIR-Arxiv	prior	0.320	**0.015 / 1.000**	1015

Table 8: WikiMIA/GPT-2 null-source diagnostics. Candidate external null pools have larger positive fractions and tail asymmetry than the evaluated WikiMIA true nulls.

Null source	Mean score	$P(W > 0)$	Tail asym. at τ_{KTD}
Evaluated WikiMIA true null	2.129	0.635	0.136
Raw Wikipedia 20231101.en	23.971	0.999	0.957
WikiMIA-25 length64	11.939	0.984	0.740
WikiMIA-25 mixed lengths	7.832	0.761	0.457

The certified rule is valid but too conservative at the available calibration-pool sizes: its best certified ratio remains above q in every setting, so no threshold is accepted. The bootstrap penalty is the practical high-power instantiation used in the main experiments. It often controls empirical FDP, but can fail under external-null mismatch, as in the Pythia/MIMIR-Arxiv rows. This separates two issues: finite-sample certification is currently power-conservative, while external-null transfer still requires the calibration and evaluated null statistics to be sufficiently matched.

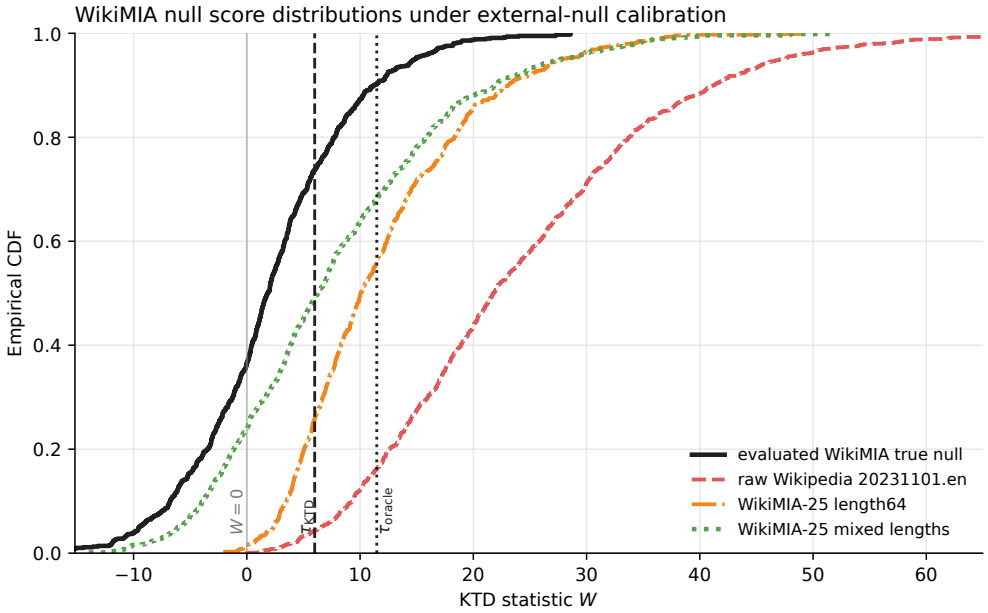

Figure 4: WikiMIA external-null stress test. Candidate external null pools are shifted toward the positive tail under the GPT-2 gradient statistic, causing AKTD (E) to overestimate the asymmetry penalty and make no discoveries at $q = 0.1$. Vertical lines mark the ordinary KTD and oracle-adjusted thresholds.

C.6 WikiMIA External-Null Stress Test

AKTD (E) makes no discoveries on WikiMIA/GPT-2 with the available external null pools. Figure 4 and Table 8 explain this negative result: relative to the evaluated WikiMIA true-null distribution, the candidate external pools are shifted toward the positive tail under the GPT-2 gradient statistic, with much larger positive fractions and tail asymmetry. AKTD (E) therefore overestimates the null-tail penalty and becomes over-conservative.

This negative case complements the successful oracle and cross-fit WikiMIA diagnostics: external-null calibration requires a matched null distribution, not merely an external non-member corpus.

D Additional Discussion and Larger-Model Results

Same-null scope and robust knockoffs. The external-null certificate requires the calibration statistics and evaluated null statistics to share a distribution after the complete knockoff-generation and scoring pipeline; genuine non-membership alone is insufficient, and observable diagnostics can flag a clear mismatch but cannot certify equality. Robust-knockoff analyses instead quantify construction error through distributional discrepancies, coupling conditions, or model error [Barber et al., 2020, Fan et al., 2025a,b]. These views are complementary: robust-knockoff theory characterizes departures from exchangeability, whereas AKTD converts the resulting realized null-tail imbalance into a pathwise FDP correction, whose guarantee requires the penalty to dominate that imbalance.

Larger-model results. We repeat the pathwise experiment on MIMIR with Pythia-6.9B, Pythia-12B, and GPT-J-6B using the same $q = 0.1$ thresholding rules.

Table 9: Larger-model pathwise results at $q = 0.1$ (empirical FDP/power). Bold FDP values meet the target.

Model	Dataset	KTD	AKTD (O)	AKTD (C)
Pythia-6.9B	MIMIR-Arxiv	0.499 / 1.000	**0.096** / 0.998	**0.068** / 0.994
Pythia-6.9B	MIMIR-PileCC	0.499 / 1.000	**0.098** / 0.944	**0.088** / 0.939
Pythia-12B	MIMIR-Arxiv	0.499 / 1.000	**0.096** / 0.998	**0.078** / 0.994
Pythia-12B	MIMIR-PileCC	0.499 / 1.000	**0.098** / 0.933	**0.085** / 0.922
GPT-J-6B	MIMIR-Arxiv	0.499 / 1.000	**0.096** / 1.000	**0.079** / 0.998
GPT-J-6B	MIMIR-PileCC	0.499 / 1.000	**0.098** / 0.988	**0.082** / 0.987

Across all six model–dataset pairs, ordinary KTD has empirical FDP 0.499, whereas both AKTD variants remain below 0.1 while retaining high power. This extends the main pathwise diagnosis without changing the procedure.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction accurately state the paper's theoretical, methodological, and empirical claims, including the limitation of external-null calibration.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses key limitations, including approximate same-null assumptions, conservative certified penalties, empirical bootstrap calibration, and limited model/dataset coverage.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All theoretical results state their assumptions and are proved in the appendix, including pathwise FDP bounds and the external-null certificate.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The paper and appendix specify datasets, models, statistics, knockoff counts, calibration protocols, bootstrap parameters, and evaluation metrics needed to reproduce the main results.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: The submission includes anonymized code and reproduction instructions for the proposed methods.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The experimental setup specifies datasets, models, statistics, hyperparameters, calibration sources, cross-fit folds, and bootstrap settings.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: The main experiments report empirical FDP and power on fixed labeled benchmark candidate sets rather than repeated random trials with error bars. We provide additional q-grid, cross-fit, estimator, and external-null diagnostics in the appendix to assess robustness, but we do not report formal confidence intervals or statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper describes the compute environment in experiment setting.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work uses public benchmarks and open model checkpoints for auditing research and is designed to reduce false positive training-data detections.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper discusses positive auditing uses for privacy, copyright, data governance, and benchmark contamination, as well as the negative risk that inaccurate training-data detections may lead to misleading accusations. The proposed method is intended to mitigate such false-positive risks.

Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release new pretrained models, scraped datasets, or other high-risk assets requiring release safeguards.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The paper cites the datasets, models, and baseline methods used in the experiments and follows their research-use settings.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

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

13. **New assets**

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

    Answer: [Yes]

    Justification: The submitted code and experiment artifacts are documented with instructions for reproducing the main results.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or human-subject experiments.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve human-subject research or collection of new personal data.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: The experiments use pretrained language models as target models and generated knockoffs as part of the evaluation pipeline.

    Guidelines:

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