CLEANREF-LFDR: FDR-CONTROLLED CONTAMINATION DETECTION WITH CLEAN REFERENCES FOR LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale pretraining has enabled large language models (LLMs) to achieve strong performance across diverse tasks. However, overlap between evaluation benchmarks and pretraining corpora introduces test data contamination, undermining evaluation reliability by confounding memorization with generalization. Existing training data detectors seek to identify clean evaluation samples, but undisclosed training corpora make direct membership verification difficult. These detectors therefore rely on indirect signals and may leave contaminated samples in the retained set. False discovery rate (FDR) control provides a principled way to limit the expected proportion of contaminated samples retained while seeking to preserve as many clean samples as possible. Yet, approaches that rely on contaminated references for calibration face a practical challenge: when the training corpus is undisclosed, the membership of these references is difficult to verify. Trusted clean references can still be available for a fixed target model, such as private data known to have been excluded from training or genuinely new content created after the model's release. This motivates us to pursue FDR control with high statistical power using only clean references. We propose CleanRef-LFDR, a sample selection framework that uses a finite trusted clean reference set and an unlabeled evaluation mixture. CleanRef-LFDR estimates the clean proportion and uses density ratios to infer local contamination risks, then selects the largest subset satisfying an average estimated-risk constraint. Accounting for proportion and density estimation errors, we establish a finite-sample FDR upper bound and asymptotic control under stated conditions. Experiments across models, datasets, and detection scores demonstrate effective empirical risk control and high statistical power in a broad range of controlled contamination settings.

1 INTRODUCTION

Large language models (LLMs) have achieved substantial progress across diverse tasks through large-scale pretraining (Brown et al., 2020; Zhao et al., 2026). However, the scale of their training data also introduces test data contamination: evaluation samples may already have appeared in the pretraining corpus (Deng et al., 2024; Jacovi et al., 2023). When a model has encountered evaluation content during training, strong test performance may reflect both memorization of training samples and generalization to unseen data, making it difficult to assess its actual capabilities (Magar & Schwartz, 2022; Dong et al., 2024; Carlini et al., 2023; Kocyigit et al., 2025). Identifying and retaining clean samples from potentially contaminated evaluation sets is therefore important for reliable model evaluation (Chang et al., 2024; Dekoninck et al., 2024).

To mitigate this issue, prior work has developed training data detection methods that exploit differences in model responses to seen and unseen samples to infer training membership (Shi et al., 2024; Zhang et al., 2025b; Xie et al., 2024). However, undisclosed training corpora make direct membership verification difficult, so these methods rely on indirect signals that may not fully distinguish contaminated from clean samples. Contaminated samples may consequently be retained as clean, compromising the reliability of model evaluation. This calls for false discovery rate (FDR) control, which constrains the expected proportion of contaminated samples in the retained set. An effective selection method should also achieve high statistical power, measured by the proportion

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.

of truly clean samples retained. FTD pursues these objectives by combining training data detection with FDR control (Zhang et al., 2026). Its calibration, however, relies on a reference distribution of contaminated scores, constructed from samples assumed to have been included in the target model's training data.

Reliable contaminated references are difficult to obtain when training corpora and data processing procedures are undisclosed: a sample's existence before training or origin from a common pretraining source does not establish its actual use in the target model's training (Duarte et al., 2024; Jacovi et al., 2023). Nevertheless, the absence of reliable contamination references does not preclude access to trusted clean data. For a fixed target model, these may include private data known to have been excluded from its training, or genuinely new content created after its release and previously nonexistent. Temporal information has also been used to construct unseen samples in prior detection studies (Shi et al., 2024; Li et al., 2024; Roberts et al., 2024). This motivates our goal: achieving FDR control with high statistical power using clean references alone.

Achieving this goal presents two connected challenges. **Methodologically**, clean reference scores characterize the clean component but do not directly specify contamination probabilities within the evaluation mixture. Without a contamination reference or a known mixture proportion, a selection framework must infer how detection scores relate to contamination risk and use that information to determine which samples to retain under a user-specified risk budget. **Theoretically**, this selection must rely on quantities estimated from finite data. Errors in proportion and density estimation can underestimate contamination risk and thereby admit too many contaminated samples. Satisfying a constraint on estimated risk therefore does not automatically establish control of the true FDR. A statistical guarantee must account for how these estimation errors propagate through the adaptive selection procedure to the final retained set.

To address these challenges, we propose **CleanRef-LFDR**, a sample selection framework that converts existing detection scores into local contamination risks using a finite clean reference and an unlabeled evaluation mixture. It first estimates the clean proportion from the relationship between the two score distributions, then combines this estimate with their density ratio to estimate each sample's contamination probability. Samples are ranked by increasing estimated risk, and the largest prefix whose average estimated risk does not exceed the target level is retained. This rule maximizes the retained set size under the estimated-risk constraint. To connect this constraint to actual residual contamination, we analyze clean-proportion overestimation and density estimation error and establish a finite-sample FDR upper bound. Under reference distribution matching, mixture identifiability, and the stated separation and regularity conditions, the error terms vanish as both the reference and mixed samples grow, yielding asymptotic FDR control at the target level. Figure 1 summarizes the framework.

We instantiate CleanRef-LFDR with multiple detection scores and evaluate it across models and datasets under controlled contamination. The experiments examine residual contamination jointly with clean-sample retention, including comparisons of score constructions and selection rules. Further analyses investigate the effects of the target risk level, reference and evaluation set sizes, and data conditions on selection performance.

Our main contributions are summarized as follows:

- **Clean-reference selection framework.** We study contamination-controlled evaluation data selection without a contamination reference and develop CleanRef-LFDR to estimate local contamination risks and select the largest feasible set under an average estimated-risk constraint.
- **Statistical guarantees.** We characterize the effects of finite-reference proportion and density estimation errors on false retention, establishing a finite-sample FDR upper bound and asymptotic control under explicit statistical conditions.
- **Empirical evaluation.** We evaluate the framework with multiple detection scores across models and datasets, examining residual contamination, clean-sample retention, and sensitivity to the available reference information and data conditions.

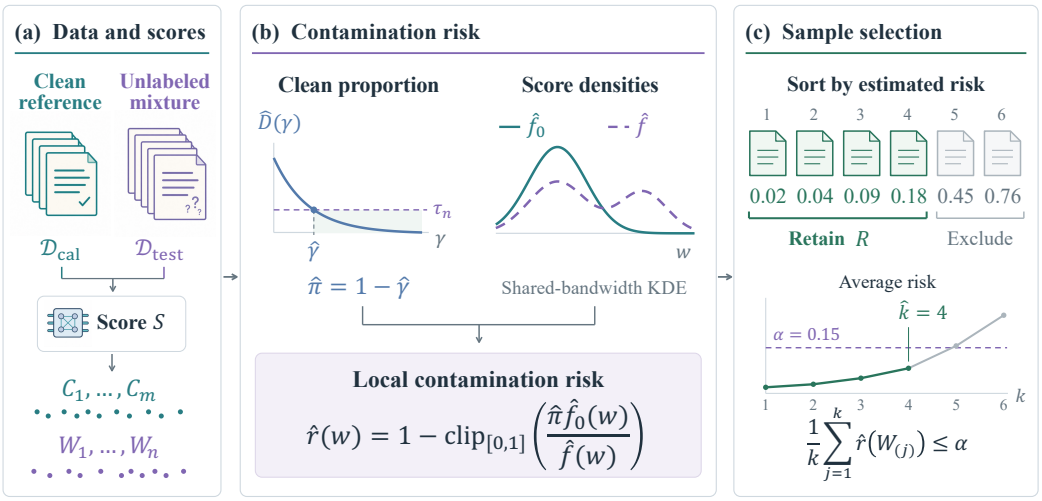

Figure 1: CleanRef-LFDR: fixed scores, clean-proportion estimation, shared-bandwidth KDE, local contamination risks, and largest-prefix selection under an average estimated-risk budget α. Curves and values are schematic; selection colors do not denote true membership.

2 Related Work

Training data detection. Training membership inference helps assess whether benchmark performance reflects memorization (Duan et al., 2024; Golchin & Surdeanu, 2024). Min-K% uses low-probability token statistics (Shi et al., 2024); Min-K%++ standardizes token log probabilities (Zhang et al., 2025b); ReCaLL uses nonmember prefixes (Xie et al., 2024); and FSD measures score changes after training on unseen data (Zhang et al., 2025a). These signals support discrimination but do not directly control contamination among retained samples. We take fixed detection scores as inputs and study risk-constrained selection.

FDR control and reference information. With contamination as the null state, a false discovery is a contaminated sample retained as clean. BH controls FDR under appropriate conditions on null p-values (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). FDR control has also been used in many machine-learning tasks (Bates et al., 2023; Jin & Candès, 2023; Marandon et al., 2024; Wang et al., 2024; Wu et al., 2024; Zhang et al., 2022b). Related guarantees address different targets: Oren et al. (2024) test dataset-level contamination, whereas KTD controls false declarations of training membership (Hu et al., 2025). Neither directly controls contamination in a retained clean subset. FTD shares our objective and combines contaminated-reference p-values through adaptive detector fusion (Zhang et al., 2026). We instead use a finite clean reference and an unlabeled mixture. In another clean-reference setting, Liu et al. (2022) bound novel-class detection rates given proportion information; that target differs from retained-set FDR. PU-BH provides a further positive–unlabeled comparison (He et al., 2026).

Our estimator builds on mixture-component projection (Patra & Sen, 2016) and local-FDR ranking and selection (Efron, 2008; Sun & Cai, 2007). We extend the finite-reference error analysis to quantify how clean-proportion overestimation and integrated density errors affect false retention, obtaining a finite-sample FDR upper bound and asymptotic control under explicit reference-matching, sampling, separation, and bandwidth conditions.

3 Problem Formulation

We use uppercase letters for random variables and lowercase letters for their realizations. For a positive integer n, let $[n] = \{1, \ldots, n\}$, and write $a \vee b = \max\{a, b\}$. Throughout the paper, clean means that a sample was not used to train the model being evaluated, whereas contaminated means that it was used for training.

Suppose the evaluator has a trusted clean reference set

$$\mathcal{D}_{\text{cal}} = \{X_j^{\text{cal}}\}_{j=1}^{m}, \qquad X_j^{\text{cal}} \overset{\text{iid}}{\sim} \mathbb{P}_0,$$

where $\mathbb{P}_0$ denotes the clean-sample distribution. The evaluator also has an unlabeled candidate evaluation set $\mathcal{D}_{\text{test}} = \{X_i\}_{i=1}^{n}$. For each candidate, let $Z_i \in \{0, 1\}$ denote its unobserved status, with $Z_i = 0$ indicating clean and $Z_i = 1$ indicating contaminated. We model

$$Z_i \overset{\text{iid}}{\sim} \text{Bernoulli}(1 - \pi), \qquad X_i \mid Z_i = z \sim \mathbb{P}_z, \quad z \in \{0, 1\}, \tag{1}$$

where $\mathbb{P}_1$ is the contaminated-sample distribution and $\pi = \Pr(Z_i = 0) \in (0, 1)$ is the population clean proportion. The true statuses and π are unavailable to the selection procedure.

Let $\mathcal{H}$ collect the fixed model state and external information used to fit the scoring function. Throughout, distributions, probabilities, and expectations are conditional on $\mathcal{H}$. Let S be a pre-specified one-dimensional scoring function and define the clean reference and candidate scores by

$$C_j = S(X_j^{\text{cal}}), \quad j \in [m], \qquad W_i = S(X_i), \quad i \in [n].$$

Let F_0 and F_1 be the cumulative distribution functions (CDFs) of scores from $\mathbb{P}_0$ and $\mathbb{P}_1$, respectively. The candidate-score distribution then satisfies the standard two-group model (Efron, 2008)

$$F = \pi F_0 + (1 - \pi) F_1. \tag{2}$$

The clean reference is therefore required not only to have trusted clean membership, but also to represent the clean component F_0 of the candidate population. A detection score may rank some samples more reliably than others, but its scale need not be probabilistically calibrated, and contamination risk need not be monotone in the raw score.

Our goal is to select a subset of candidate samples, indexed by $R \subseteq [n]$, while controlling the expected proportion of contaminated samples in the retained set at a user-specified level $\alpha \in (0, 1)$. Specifically,

$$\text{FDR}(R) = \mathbb{E}\left[\frac{\sum_{i \in R} Z_i}{|R| \vee 1}\right] \leq \alpha. \tag{3}$$

The ratio inside the expectation is the realized false discovery proportion (FDP). The ratio is defined as zero when R is empty. To rule out the trivial solution $R = \varnothing$, we also evaluate statistical power, defined as the expected proportion of clean candidate samples that are retained:

$$\text{Power} = \mathbb{E}\left[\frac{\sum_{i \in R}(1 - Z_i)}{(\sum_{i=1}^{n}(1 - Z_i)) \vee 1}\right]. \tag{4}$$

We seek high Power subject to Eq. (3). Since labels are unavailable, Section 4 maximizes retained-set size under an estimated-risk constraint; Section 5 relates it to true FDR.

4 METHOD

CleanRef-LFDR estimates local contamination risks from a clean reference and an unlabeled mixture, then selects the largest set satisfying an average estimated-risk constraint.

4.1 FROM DETECTION SCORES TO LOCAL CONTAMINATION RISK

Let f_0, f_1, and f denote the densities associated with F_0, F_1, and F. By Eq. (2), $f = \pi f_0 + (1-\pi) f_1$. For any w with $f(w) > 0$, Bayes' rule gives

$$r(w) = \Pr(Z = 1 \mid W = w) = 1 - \frac{\pi f_0(w)}{f(w)}. \tag{5}$$

We call $r(w)$ the *local contamination risk*. Under the convention that contamination is the null state, it is also the local false discovery rate (Efron, 2008).

This quantity connects individual scores to retained-set risk. Under the sampling condition in Assumption 1, any set R determined from the observed reference and candidate scores satisfies

$$\mathbb{E}\left[\frac{\sum_{i \in R} Z_i}{|R| \vee 1} \,\middle|\, \mathcal{D}_{\text{cal}}, W_{1:n}\right] = \frac{\sum_{i \in R} r(W_i)}{|R| \vee 1}. \tag{6}$$

Thus, average true risk determines conditional expected FDP. We estimate it using only the clean reference and unlabeled mixture.

4.2 Estimating the Local FDR

Clean-proportion estimation. Write $\gamma = 1-\pi$ for the contamination proportion. For a candidate value $\gamma \in (0,1]$, rearranging the mixture model gives the residual distribution

$$G_\gamma = \frac{F-(1-\gamma)F_0}{\gamma}. \tag{7}$$

A compatible G_γ must be a CDF. Too small a contamination share can violate this condition, motivating the smallest compatible proportion via the projection approach of Patra & Sen (2016).

For finite samples, let $x_1 \le \cdots \le x_n$ be the ordered candidate scores and let $\widehat{F}_0$ be the empirical CDF of the reference scores $C_1,\dots,C_m$. For a candidate γ, define its empirical incompatibility by

$$\widehat{D}(\gamma) = \min_{0\le g_1\le\cdots\le g_n\le 1} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left[\frac{i}{n} - (1-\gamma)\widehat{F}_0(x_i) - \gamma g_i\right]^2 \right\}^{1/2}. \tag{8}$$

This isotonic projection measures compatibility with a valid residual CDF.

We use a positive tolerance τ_n that decreases with candidate sample size and define the continuous-parameter estimator

$$\widehat{\gamma} = \inf\{\gamma \in (0,1] : \widehat{D}(\gamma) \le \tau_n\}, \qquad \widehat{\pi} = 1-\widehat{\gamma}. \tag{9}$$

The distance extends continuously to $\gamma = 0$, so the infimum may be zero. Since $\widehat{D}$ is nonincreasing, the implementation locates the feasible boundary by bisection to interval width 10^{-8}. The tolerance and numerical safeguards are specified in Appendix B; the theory analyzes the exact statistical estimator. This selects the largest empirically compatible clean proportion. Overestimating π inflates the clean posterior and underestimates risk; Section 5 controls this one-sided error.

Density estimation. After estimating the clean proportion, we construct kernel density estimates (KDEs) (Parzen, 1962) for the clean and mixture densities using a common effective bandwidth h and the Gaussian kernel $K(u) = (2\pi)^{-1/2}e^{-u^2/2}$:

$$\widehat{f}_0(w) = \frac{1}{mh}\sum_{j=1}^{m} K\left(\frac{w-C_j}{h}\right), \qquad \widehat{f}(w) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{w-W_i}{h}\right). \tag{10}$$

The bandwidth and numerical safeguards used in the experiments are specified in Appendix B.3.

Local-FDR construction. The estimated clean posterior and local contamination risk are

$$\widehat{\ell}(w) = \mathrm{clip}_{[0,1]}\left(\frac{\widehat{\pi}\widehat{f}_0(w)}{\widehat{f}(w)}\right), \qquad \widehat{r}(w) = 1-\widehat{\ell}(w), \tag{11}$$

Here, $\mathrm{clip}_{[0,1]}(u) = \min\{1, \max\{0,u\}\}$.

4.3 Risk-Constrained Sample Selection

We maximize the number retained under an average estimated-risk constraint:

$$\max_{R\subseteq[n]} |R| \quad \text{subject to} \quad \frac{\sum_{i\in R}\widehat{r}(W_i)}{|R|\vee 1} \le \alpha. \tag{12}$$

Sort candidates by increasing estimated risk, with (j) denoting the original index at rank j, and retain the longest feasible prefix (Sun & Cai, 2007):

$$\widehat{k} = \max\left(\{0\}\cup\left\{k\in[n] : \frac{1}{k}\sum_{j=1}^{k}\widehat{r}(W_{(j)}) \le \alpha\right\}\right), \qquad R = \{(1),\dots,(\widehat{k})\}. \tag{13}$$

For a fixed size k, this prefix has the smallest total estimated risk. If it is infeasible, every size-k subset is infeasible. Since prefix-average risk is nondecreasing, Eq. (13) solves Eq. (12). Section 5 links this estimated-risk constraint to true FDR.

5 THEORETICAL GUARANTEES

We establish FDR control when the clean score distribution is estimated from a finite reference sample. We first extend the Patra–Sen projection analysis to obtain a finite-sample bound on clean-proportion overestimation, accounting for uncertainty in both empirical distributions. We then propagate this error and density estimation error to the retained set, yielding a finite-sample FDR bound and asymptotic control. Throughout, probabilities are conditional on $\mathcal{H}$, and $N = \min(m, n) \geq 3$. Proofs are given in Appendix A.

Assumption 1 (Conditional sampling)**.** *Conditional on $\mathcal{H}$, the reference scores are IID from F_0, the pairs $(W_i, Z_i)_{i=1}^n$ are IID from Eq.* (2)*, and the two samples are independent.*

Thus, the reference must represent the clean component of the evaluation population; trusted clean membership alone does not ensure distributional matching.

Assumption 2 (Identifiability and quantitative separation)**.** *F_0, F are continuous, and F_1 cannot be written as $\lambda F_0 + (1-\lambda)G$ for any $\lambda \in (0, 1]$ and CDF G. Define*

$$D(\gamma) = \inf_{G \text{ a CDF}} \left\{ \int \left[F - (1-\gamma)F_0 - \gamma G\right]^2 dF \right\}^{1/2}. \tag{14}$$

For some $\kappa > 0$ and $0 < t_0 < 1 - \pi$,

$$D(1 - \pi - t) \geq \kappa t, \qquad 0 < t \leq t_0. \tag{15}$$

The first condition is the standard irreducibility condition: the contaminated component contains no positive mixture share distributed exactly as F_0. For densities, it is equivalent to $\operatorname{ess\,inf}_{w \sim F_0} f_1(w)/f_0(w) = 0$ and identifies the clean proportion (Patra & Sen, 2016; Zhu et al., 2023). The linear separation condition strengthens identifiability to obtain an explicit finite-sample rate. A sufficient condition is an interval I with $F_0(I) > 0$ and $F_1(I) = 0$, an interval version of the standard anchor-set conditions used in mixture proportion estimation (Ramaswamy et al., 2016; Garg et al., 2021).

Theorem 1 (Finite-reference Patra–Sen bound)**.** *Under Assumptions 1 and 2, for every $\delta \in (0, 1)$, the estimator in Eq.* (9) *satisfies, with probability at least $1 - \delta$,*

$$(\widehat{\pi} - \pi)_+ \leq \frac{1}{\kappa} \left\{ \tau_n + \sqrt{\frac{\log(4/\delta)}{2m}} + \sqrt{\frac{\log(4/\delta)}{2n}} + \frac{1}{n} \right\}. \tag{16}$$

Here $(x)_+ = \max(x, 0)$.

The $m^{-1/2}$ and $n^{-1/2}$ terms quantify uncertainty from the reference and mixed samples, respectively. For the tolerance in Eq. (35) and $\delta = 1/N$, Theorem 1 gives $(\widehat{\pi} - \pi)_+ \leq 4\kappa^{-1}\sqrt{\log N/N}$ with probability at least $1 - 1/N$. Overestimation of the clean proportion is the direction that can underestimate contamination risk. While Patra & Sen (2016) establish convergence rates and confidence bounds when F_0 is known, Theorem 1 extends the projection analysis to an estimated reference distribution. The resulting explicit finite-sample error bound separates the contributions of the two samples and supports the subsequent FDR analysis.

Next, we impose the following regularity conditions for kernel density estimation (Tsybakov & Zaiats, 2009; Huo et al., 2024).

Assumption 3 (Integrated density regularity and bandwidth)**.** *For $g = f_0, f$, the density g is twice continuously differentiable and satisfies*

$$\int_{\mathbb{R}} |g''(w)|\, dw \leq L, \qquad \int_{\mathbb{R}} w^2 g(w)\, dw \leq V.$$

The Gaussian KDEs use a common bandwidth between uh and vh, where $h \in [N^{-1}, 1]$ is deterministic and $0 < u \leq v$. The bandwidth may depend on the clean reference, but not on the mixed sample. The constants L, V, u, v are fixed conditional on $\mathcal{H}$ and independent of m, n.

These conditions give expected integrated KDE error of standard order $h^2 + (Nh)^{-1/2}$.

Theorem 2 (Finite-sample FDR bound). *Under Assumptions 1–3, with a positive constant $C > 0$, the set returned by Eq.* (13) *satisfies*

$$\mathrm{FDR}(R) \le \alpha + C\Bigg\{ \underbrace{\frac{\tau_n + N^{-1/2}}{\kappa}}_{\text{proportion estimation}} + \underbrace{h^2}_{\text{KDE bias}} + \underbrace{\frac{1}{\sqrt{Nh}}}_{\text{KDE fluctuation}} + \underbrace{\exp\!\left(-\frac{N\tau_n^2}{8}\right)}_{\text{projection feasibility}} \Bigg\}. \tag{17}$$

If $h \to 0$, $Nh \to \infty$, $\tau_n \to 0$, and $N\tau_n^2 \to \infty$, then

$$\limsup_{m,n\to\infty} \mathrm{FDR}(R) \le \alpha. \tag{18}$$

The first three terms separate proportion uncertainty from density smoothing and sampling error. The last term controls the probability that the true contamination proportion fails the empirical projection criterion. For $h = N^{-1/5}$, the KDE terms are $O(N^{-2/5})$. With the original tolerance in Eq. (35), $m \asymp n$ ensures $N\tau_n^2 \to \infty$. The proof uses average risk error and a leave-one-out comparison, permitting the mixed samples to be used for both KDE and selection.

6 Experiments

This section presents experimental results to evaluate the effectiveness of CleanRef-LFDR.[1]

6.1 Experimental Setup

Models and datasets. We evaluate CleanRef-LFDR under controlled contamination using Pythia-2.8B (Biderman et al., 2023), OPT-6.7B (Zhang et al., 2022a), and LLaMA-13B (Touvron et al., 2023) on WikiMIA (Shi et al., 2024), arXivTection (Duarte et al., 2024), and BBC Real Time (news) (Li et al., 2024). Dataset descriptions and the experimental construction are provided in Appendix B.1.

Scores and baselines. We compare CleanRef-LFDR with PU-BH (He et al., 2026), which selects samples from mixed data using clean references. We apply PU-BH to the original and fsd scores, yielding two baselines denoted by Raw and fsd, respectively. The experiments include nine scoring configurations covering likelihood, gradient, Min-K, PAC, and combined scores; their definitions are provided in Appendix B.2. All methods use the same clean reference set, candidate evaluation set, and target risk level.

Evaluation metrics. We jointly evaluate empirical FDR and statistical power (Power). Empirical FDR is the mean FDP over repeated trials, and Power is the mean proportion of clean samples retained in each trial. The main experiments are repeated 100 times, and we report the average results. Further details are provided in Appendix B.3.

6.2 Main Results

We set the target FDR level to $\alpha = 0.15$ and compare the methods across three datasets, three models, and nine scoring configurations. Table 1 reports the complete results. The main observations are as follows.

Broad empirical risk control. CleanRef-LFDR keeps empirical FDR below the target level in most configurations, including all evaluated configurations on WikiMIA and arXivTection, demonstrating effective contamination control across models and scoring configurations.

Effective recovery of clean evaluation samples. CleanRef-LFDR achieves higher power than both baselines in most configurations. It attains near-perfect power in most configurations on WikiMIA and arXivTection and maintains high clean-sample retention across multiple news configurations. These results show that CleanRef-LFDR can limit contamination while preserving the clean data needed for evaluation.

[1]The code is publicly available at `https://anonymous.4open.science/r/cleanref-lfdr-5411/`.

Table 1: Main results at $\alpha = 0.15$. Entries are empirical FDR / Power over 100 repeated experiments. Bold marks the highest Power among methods meeting the target.

	WikiMIA						arXivTection						news					
	Raw		fsd		**Ours**		Raw		fsd		**Ours**		Raw		fsd		**Ours**	
Score	FDR↓	Power↑	FDR↓	Power↑	FDR↓	Power↑	FDR↓	Power↑	FDR↓	Power↑	FDR↓	Power↑	FDR↓	Power↑	FDR↓	Power↑	FDR↓	Power↑
Pythia-2.8B																		
PPL	0.193	0.870	0.044	0.889	**0.055**	**0.990**	0.321	0.784	0.042	0.866	**0.104**	**0.992**	0.334	0.850	0.365	0.829	**0.018**	**0.598**
Lowercase	0.267	0.846	0.046	0.875	**0.081**	**0.986**	0.401	0.720	0.200	0.836	**0.130**	**0.975**	0.352	0.851	0.262	0.876	**0.141**	**0.903**
Zlib	0.411	0.733	0.042	0.869	**0.069**	**0.988**	0.414	0.634	0.018	0.885	**0.105**	**0.995**	0.385	0.829	0.229	0.882	**0.085**	**0.916**
Grad	0.405	0.735	0.195	0.855	**0.038**	**0.851**	0.369	0.746	0.093	0.859	**0.072**	**0.983**	0.378	0.823	0.283	0.880	**0.112**	**0.877**
Min-K 20%	0.162	0.864	0.007	0.895	**0.064**	**0.994**	0.236	0.850	0.021	0.849	**0.091**	**1.000**	0.358	0.839	0.283	0.874	**0.147**	**0.919**
Min-K 30%	0.161	0.850	0.004	0.885	**0.047**	**0.989**	0.238	0.828	0.027	0.889	**0.093**	**0.999**	0.354	0.848	0.242	0.891	0.156	0.960
PAC	0.272	0.812	0.093	0.854	**0.122**	**1.000**	0.357	0.766	0.097	0.839	**0.105**	**0.991**	0.397	0.769	0.312	0.831	0.154	0.911
MEAN_Z	0.146	0.871	0.000	0.899	**0.051**	**0.989**	0.257	0.846	0.003	0.873	**0.104**	**1.000**	0.328	0.863	0.210	0.904	**0.149**	**0.970**
First_PCA	0.144	0.872	0.000	0.879	**0.053**	**0.989**	0.232	0.837	0.004	0.878	**0.096**	**1.000**	0.331	0.841	0.207	0.884	0.151	0.971
OPT-6.7B																		
PPL	0.346	0.814	0.058	0.864	**0.083**	**0.979**	0.374	0.739	0.097	0.869	**0.065**	**0.973**	0.389	0.774	0.367	0.833	**0.065**	**0.670**
Lowercase	0.421	0.721	0.148	0.875	**0.092**	**0.943**	0.428	0.649	0.224	0.854	**0.107**	**0.924**	0.392	0.798	0.323	0.851	0.163	0.846
Zlib	0.456	0.635	0.057	0.858	**0.119**	**0.990**	0.458	0.575	0.037	0.867	**0.048**	**0.966**	0.414	0.727	0.261	0.879	**0.109**	**0.900**
Grad	0.438	0.704	0.212	0.836	**0.063**	**0.887**	0.392	0.717	0.217	0.835	**0.043**	**0.766**	0.391	0.797	0.362	0.835	**0.083**	**0.626**
Min-K 20%	0.293	0.815	0.019	0.867	**0.095**	**1.000**	0.308	0.816	0.065	0.888	**0.110**	**1.000**	0.399	0.767	0.335	0.863	**0.137**	**0.825**
Min-K 30%	0.298	0.845	0.007	0.883	**0.091**	**1.000**	0.340	0.792	0.020	0.878	**0.099**	**0.997**	0.389	0.805	0.304	0.863	**0.115**	**0.876**
PAC	0.363	0.800	0.079	0.859	**0.094**	**0.984**	0.403	0.706	0.103	0.855	**0.088**	**0.968**	0.445	0.682	0.373	0.804	**0.116**	**0.742**
MEAN_Z	0.277	0.847	0.008	0.893	**0.087**	**1.000**	0.308	0.787	0.014	0.879	**0.079**	**0.993**	0.378	0.826	0.238	0.873	**0.120**	**0.922**
First_PCA	0.285	0.825	0.007	0.888	**0.087**	**1.000**	0.304	0.791	0.014	0.865	**0.099**	**0.998**	0.381	0.817	0.231	0.879	**0.123**	**0.923**
LLaMA-13B																		
PPL	0.036	0.843	0.010	0.880	**0.060**	**0.989**	0.226	0.843	0.023	0.887	**0.111**	**0.999**	0.214	0.882	0.309	0.851	**0.069**	**0.835**
Lowercase	0.069	0.891	0.027	0.885	**0.075**	**0.988**	0.328	0.798	0.194	0.859	**0.098**	**0.962**	0.236	0.882	0.168	0.895	0.151	0.983
Zlib	0.213	0.846	0.001	0.885	**0.115**	**0.993**	0.350	0.778	0.000	0.885	**0.072**	**0.997**	0.296	0.857	0.147	0.891	**0.121**	**0.979**
Grad	0.291	0.836	0.150	0.843	**0.040**	**0.893**	0.269	0.826	0.154	0.851	**0.088**	**0.968**	0.310	0.861	0.218	0.882	**0.135**	**0.942**
Min-K 20%	0.026	0.893	0.007	0.889	**0.040**	**0.983**	0.082	0.857	0.000	0.882	**0.086**	**0.998**	0.223	0.874	0.163	0.880	**0.134**	**0.979**
Min-K 30%	0.027	0.889	0.002	0.877	**0.055**	**0.991**	0.134	0.853	0.000	0.882	**0.086**	**0.998**	0.216	0.887	0.142	0.883	**0.131**	**0.990**
PAC	0.061	0.881	0.000	0.884	**0.102**	**1.000**	0.175	0.809	0.030	0.868	**0.079**	**0.994**	0.310	0.826	0.222	0.854	**0.116**	**0.964**
MEAN_Z	0.006	0.872	0.000	0.913	**0.091**	**1.000**	0.140	0.857	0.006	0.878	**0.076**	**1.000**	0.187	0.889	0.089	0.894	**0.148**	**0.995**
First_PCA	0.017	0.925	0.000	0.880	**0.093**	**0.997**	0.138	0.864	0.000	0.890	**0.095**	**1.000**	0.201	0.891	0.092	0.900	**0.144**	**0.994**

Higher power under the same risk requirement. In every configuration where both CleanRef-LFDR and fsd meet the risk target, CleanRef-LFDR achieves higher power. This shows that the method can retain more clean samples while satisfying the same risk requirement. Appendix E further examines the roles of proportion and density estimation in this process.

6.3 Analysis of CleanRef-LFDR

Effect of the target risk level. We vary $\alpha \in \{0.05, 0.10, 0.15, 0.18, 0.20\}$. Across all three datasets, empirical FDR remains below the corresponding target while power approaches one as the budget increases (Figure 2). Nearly all clean samples are retained at the default level, supporting effective adaptation to the risk budget.

Effect of clean-reference size. With the candidate set fixed and $\alpha = 0.15$, we increase the reference size from 25% to 100% of the full set. Empirical FDR stays below the target and power remains high on all three datasets (Figure 3). Retaining the vast majority of clean samples with one quarter of the references supports effective selection with limited reference data in these settings.

Additional analyses of reference-domain composition are provided in Appendix C.4.

6.4 Application of CleanRef-LFDR

Following FTD (Zhang et al., 2026), we evaluate eight models on SimpleQA (Wei et al., 2024) and ARC-C (Clark et al., 2018). We introduce controlled contamination (Hu et al., 2025) by fine-tuning Mistral-7B-Instruct-v0.3 on designated questions, with contaminated questions comprising 50% of

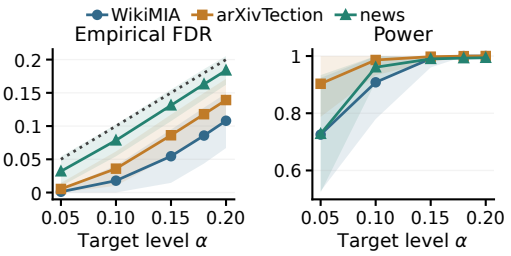

Figure 2: Target-level analysis (LLaMA-13B, Min-K 30%). Dataset-wise means with one-standard-deviation bands; dotted line: $y = \alpha$. Full results: Appendix C.1.

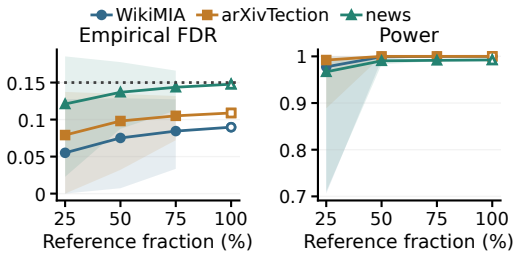

Figure 3: Effect of clean-reference size (LLaMA-13B, Min-K 30%, $\alpha = 0.15$). Bands show the 2.5th–97.5th percentiles; the full-reference endpoint has no band.

the candidate set. All models are re-evaluated on each method's retained subset; details appear in Appendix F.

Because methods retain different question sets, we compare rankings rather than absolute accuracies (Sun et al., 2025). Spearman's rank correlation coefficient (SRCC) (Sedgwick, 2014) measures agreement with the pre-injection ranking on the unfiltered candidate set; values closer to one indicate greater consistency. Table 2 uses Min-K 30% at $\alpha = 0.15$.

Table 2: Selection results using Min-K 30% at $\alpha = 0.15$: realized FDP, Power, and SRCC relative to the pre-injection ranking on the full candidate set. Unfiltered reports SRCC after injection and before selection.

	SimpleQA			ARC-C		
Method	FDP↓	Power↑	SRCC↑	FDP↓	Power↑	SRCC↑
Unfiltered	—	—	0.707	—	—	0.929
Raw	0.286	0.765	0.952	0.098	0.877	0.952
fsd	0.188	0.824	1.000	0.038	0.874	0.946
CleanRef-LFDR	0.142	0.945	1.000	0.107	1.000	0.952

Contamination changes the pre-injection ordering on both benchmarks, illustrating how exposure can distort model comparisons. CleanRef-LFDR recovers the complete pre-injection ordering on SimpleQA and improves consistency on ARC-C, matching the best baseline SRCC on both benchmarks. On both benchmarks, CleanRef-LFDR keeps FDP below the target and achieves higher power than both baselines, attaining this ranking consistency while retaining more clean questions for evaluation.

7 Conclusion and Limitations

CleanRef-LFDR uses a finite clean reference set and an unlabeled mixture to estimate contamination risks and select evaluation data. Accounting for proportion-estimation error and integrated density-estimation error, we establish a finite-sample FDR upper bound and asymptotic control under the stated conditions. Experiments demonstrate empirical FDR control across multiple configurations and greater clean-sample retention under the same risk requirement. The application study further shows improved ranking consistency and use of clean evaluation data. Our experiments mainly use controlled contamination with known training-exposure labels, which differs from complex, difficult-to-verify natural pretraining exposure. Further evaluation should cover natural contamination and weak exposure across training histories and independent data sources. Clean references may also inadequately represent the clean component of the candidate set. Although we analyze reference size and source, complex distribution shifts may still affect risk estimation and selection (Tibshirani et al., 2019; Hu et al., 2026). Maintaining reliable risk control and high clean-sample retention with limited or shifted reference information requires study.

AI USE STATEMENT

We used generative AI tools to aid or polish writing, including improving clarity, wording, and readability. The authors made the final editorial decisions and take full responsibility for the accuracy and integrity of the manuscript, including its scientific claims, references, and conclusions.

REFERENCES

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149 – 178, 2023. doi: 10.1214/22-AOS2244. URL https://doi.org/10.1214/22-AOS2244.

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 (Methodological)*, 57(1):289–300, 01 1995. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165 – 1188, 2001. doi: 10.1214/aos/1013699998. URL https://doi.org/10.1214/aos/1013699998.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Sebastian Bordt, Suraj Srinivas, Valentyn Boreiko, and Ulrike Von Luxburg. How much can we forget about data contamination? In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4998–5016. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/bordt25a.html.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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)*, pp. 2633–2650, 2021.

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*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Jasper Dekoninck, Mark N Müller, and Martin Vechev. Constat: Performance-based contamination detection in large language models. *Advances in Neural Information Processing Systems*, 37: 92420–92464, 2024.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8706–8719, 2024.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12039–12050, 2024.

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? In *Conference on Language Modeling (COLM)*, 2024.

André Vicente Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. DE-COP: Detecting copyrighted content in language models training data. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11940–11956. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/duarte24a.html.

Bradley Efron. Microarrays, Empirical Bayes and the Two-Groups Model. *Statistical Science*, 23(1): 1 – 22, 2008. doi: 10.1214/07-STS236. URL https://doi.org/10.1214/07-STS236.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Saurabh Garg, Yifan Wu, Alexander Smola, Sivaraman Balakrishnan, and Zachary Lipton. Mixture proportion estimation and pu learning:a modern approach. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8532–8544. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/47b4f1bfdf6d298682e610ad74b37dca-Paper.pdf.

Shahriar Golchin and Mihai Surdeanu. Time travel in llms: Tracing data contamination in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 43008–43029, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *nature*, 585(7825):357–362, 2020.

Siying He, Weijuan Liang, and Jiatong Liu. Identifying labeling mechanism in positive–unlabeled learning under unknown class prior. In Emilija Perković and Daniel Malinsky (eds.), *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 2119–2136. PMLR, 17–21 Aug 2026. URL https://proceedings.mlr.press/v337/he26a.html.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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. URL https://openreview.net/forum?id=XAN8G0rvoB.

Zirui Hu, Zheng Zhang, Yingjie Wang, Leszek Rutkowski, and Dacheng Tao. Cofact: Conformal factuality guarantees for language models under covariate shift. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 116267–116297, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/bd074b828bbd782d4bece89306caff84-Paper-Conference.pdf.

Yuyang Huo, Lin Lu, Haojie Ren, and Changliang Zou. Real-time selection under general constraints via predictive inference. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 61267–61305. Curran Associates, Inc., 2024. doi: 10.52202/079017-1959. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/70b8505ac79e3e131756f793cd80eb8d-Paper-Conference.pdf.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5075–5084, 2023.

Matthew Jagielski, Om Thakkar, Florian Tramer, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring forgetting of memorized training examples. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=7bJizxLKrR.

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

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065):20150202, 2016.

Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. Overestimation in LLM evaluation: A controlled large-scale study on data contamination's impact on machine translation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31105–31132. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/kocyigit25a.html.

Yucheng Li. Estimating contamination via perplexity: Quantifying memorisation in language model evaluation. *arXiv preprint arXiv:2309.10677*, 2023.

Yucheng Li, Frank Guerin, and Chenghua Lin. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18600–18607, 2024.

Si Liu, Risheek Garrepalli, Dan Hendrycks, Alan Fern, Debashis Mondal, and Thomas G Dietterich. Pac guarantees and effective algorithms for detecting novel categories. *Journal of Machine Learning Research*, 23(44):1–47, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 157–165, 2022.

Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157–183, 2024.

P. Massart. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269 – 1283, 1990. doi: 10.1214/aop/1176990746. URL https://doi.org/10.1214/aop/1176990746.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KS8mIvetg2.

Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Rohit Kumar Patra and Bodhisattva Sen. Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(4):869–893, 2016.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International conference on machine learning*, pp. 2052–2060. PMLR, 2016.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. To the cutoff... and beyond? a longitudinal perspective on llm data contamination. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 10795–10862, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/2d04d97593c8c33d415337f408ed0e1b-Paper-Conference.pdf.

Philip Sedgwick. Spearman's rank correlation coefficient. *BMJ*, 349, 2014. ISSN 0959-8138. doi: 10.1136/bmj.g7327. URL https://www.bmj.com/content/349/bmj.g7327.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, volume 2024, pp. 51826–51843, 2024.

Wenguang Sun and T Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007.

Yifan Sun, Han Wang, Dongbai Li, Gang Wang, and Huan Zhang. The emperor's new clothes in benchmarking? A rigorous examination of mitigation strategies for LLM benchmark data contamination. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57728–57753. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/sun25t.html.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Xiaoning Wang, Yuyang Huo, Liuhua Peng, and Changliang Zou. Conformalized multiple testing after data-dependent selection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=8wvH0RZPsG.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Xiaoyang Wu, Yuyang Huo, Haojie Ren, and Changliang Zou. Optimal subsampling via predictive inference. *Journal of the American Statistical Association*, 119(548):2844–2856, 2024.

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 Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8671–8689, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.493. URL https://aclanthology.org/2024.emnlp-main.493/.

Wentao Ye, Jiaqi Hu, Liyao Li, Haobo Wang, Gang Chen, and Junbo Zhao. Data contamination calibration for black-box LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10845–10861, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.644. URL https://aclanthology.org/2024.findings-acl.644/.

Hengxiang Zhang, Songxin Zhang, Bingyi Jing, and Hongxin Wei. Fine-tuning can help detect pretraining data from large language models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=X8dzvdkQwO.

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*, 2025b. URL https://openreview.net/forum?id=ZGkfoufDaU.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022a.

Yifan Zhang, Haiyan Jiang, Haojie Ren, Changliang Zou, and Dejing Dou. Automs: Automatic model selection for novelty detection with error rate control. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 19917–19929. Curran Associates, Inc., 2022b. doi: 10.52202/068431-1448. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7dced224614f3c50b1626473f48312bf-Paper-Conference.pdf.

Zheng Zhang, Qi Liu, Siyuan Liang, Ning Li, Zirui Hu, Weibo Gao, Rui Li, Zhenya Huang, Leszek Rutkowski, Baosheng Yu, et al. Controllable contamination detection for reliable llm evaluation with statistical guarantees. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30122–30143, 2026.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, et al. A survey of large language models. *Frontiers of Computer Science*, 20(12):2012627, 2026.

Yilun Zhu, Aaron Fjeldsted, Darren Holland, George Landon, Azaree Lintereur, and Clayton Scott. Mixture proportion estimation beyond irreducibility. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42962–42982. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhu23c.html.

A PROOFS OF THE THEORETICAL GUARANTEES

All probabilities and expectations in this appendix are conditional on $\mathcal{H}$. We fix $0 < \pi, \alpha < 1$, write $\eta = 1-\pi$, and let $C_1, \dots, C_m$ and $W_1, \dots, W_n$ denote the reference and mixed scores, respectively. The empirical CDFs are $\widehat{F}_0$ and $\widehat{F}$. For a CDF H, write $\|a\|_H = (\int a^2\, dH)^{1/2}$; the notation $\|a\|_1$ denotes integration of $|a|$ with respect to Lebesgue measure.

This appendix analyzes the exact continuous-parameter estimator in Eq. (9); the finite bisection tolerance is a numerical implementation detail. The empirical projection distance and estimator are

$$\widehat{D}(\gamma) = \inf_{G \text{ a CDF}} \|\widehat{F} - (1-\gamma)\widehat{F}_0 - \gamma G\|_{\widehat{F}}, \qquad 0 \le \gamma \le 1, \tag{19}$$

$$\widehat{\gamma} = \inf\{\gamma \in (0,1] : \widehat{D}(\gamma) \le \tau_n\}, \qquad \widehat{\pi} = 1 - \widehat{\gamma}. \tag{20}$$

At $\gamma = 0$ the objective is independent of G. Both D and $\widehat{D}$ extend to $[0,1]$ as 1-Lipschitz functions: for fixed G, changing γ changes the residual norm by at most $|\gamma - \gamma'|$. Since $\widehat{D}(1) = 0$, the infimum defining $\widehat{\gamma}$ is well defined, may equal zero, and satisfies $\widehat{D}(\widehat{\gamma}) \le \tau_n$.

Let K be the standard Gaussian density, $K_s(x) = s^{-1}K(x/s)$, and let $\sigma \in [uh, vh]$ denote the actual common bandwidth. Under Assumption 3, σ is measurable with respect to the reference sample and $\mathcal{H}$. The KDEs are

$$\widehat{f}_0(w) = \frac{1}{m}\sum_{j=1}^{m} K_\sigma(w - C_j), \qquad \widehat{f}(w) = \frac{1}{n}\sum_{i=1}^{n} K_\sigma(w - W_i).$$

Write

$$r(w) = 1 - \frac{\pi f_0(w)}{f(w)}, \qquad \widehat{r}(w) = 1 - \min\left\{\frac{\widehat{\pi}\widehat{f}_0(w)}{\widehat{f}(w)}, 1\right\}.$$

The true risk may be assigned an arbitrary value on $\{f = 0\}$, an F-null set. Positive Gaussian KDEs make the estimated ratio well defined. Constants denoted by C may change between displays; their dependence is specified below.

A.1 UNIFORM COMPARISON OF PROJECTION DISTANCES

Lemma 1 (Finite-reference projection error)**.** *Under Assumption 1 and continuity of F, F_0, almost surely,*

$$\sup_{\gamma\in[0,1]} |\widehat{D}(\gamma) - D(\gamma)| \le \|\widehat{F} - F\|_\infty + \|\widehat{F}_0 - F_0\|_\infty + \frac{1}{n}. \tag{21}$$

Proof. We first compare objectives with an arbitrary continuous CDF H. For $U = F - H$, the Stieltjes chain rule gives

$$\int U^2\, dF - \int U^2\, dH = \int U^2\, dU = \tfrac{1}{3}[U^3]_{-\infty}^{\infty} = 0,$$

so $\|F - H\|_F = \|F - H\|_H$. The mixed observations are distinct almost surely by continuity of F. For their ordered values $x_1 < \cdots < x_n$, set $v_i = i/n - H(x_i)$. Integration between consecutive values of $H(x_i)$ yields the Cramér–von Mises identity

$$\|\widehat{F} - H\|_H^2 = \frac{1}{n}\sum_{i=1}^{n}\left(v_i - \frac{1}{2n}\right)^2 + \frac{1}{12n^2}, \qquad \|\widehat{F} - H\|_{\widehat{F}}^2 = \frac{1}{n}\sum_{i=1}^{n} v_i^2.$$

Viewing these as squared norms of augmented Euclidean vectors gives

$$\left|\|\widehat{F} - H\|_H - \|\widehat{F} - H\|_{\widehat{F}}\right| \le \sqrt{\frac{1}{4n^2} + \frac{1}{12n^2}} \le \frac{1}{n}.$$

The reverse triangle inequality in $L^2(H)$ therefore implies, simultaneously for continuous H,

$$\left|\|F - H\|_F - \|\widehat{F} - H\|_{\widehat{F}}\right| \le \|\widehat{F} - F\|_\infty + \frac{1}{n}.$$

Apply this with $H = (1-\gamma)F_0 + \gamma G$ for continuous G. Replacing F_0 by $\widehat{F}_0$ changes the empirical objective by at most $(1-\gamma)\|\widehat{F}_0 - F_0\|_\infty$. At the population level, continuous CDFs are dense among all CDFs in $L^2(F)$ because F is continuous. At the empirical level, every vector $0 \le G(x_1) \le \cdots \le G(x_n) \le 1$ can be interpolated by a continuous CDF. Thus restricting either infimum to continuous G does not change it. Taking infima establishes the bound for each γ, with a right-hand side independent of γ. □

A.2 Proof of the finite-reference Patra–Sen bound

Proof of Theorem 1. The population distance D is convex. Indeed, for $\lambda \in [0,1]$ and $\gamma_\lambda = \lambda\gamma_1 + (1-\lambda)\gamma_2 > 0$, the CDF

$$G_\lambda = \frac{\lambda\gamma_1 G_1 + (1-\lambda)\gamma_2 G_2}{\gamma_\lambda}$$

makes its residual equal to the corresponding convex combination of residuals. The triangle inequality and approximate minimization prove convexity; the endpoint follows by continuity. Also $D(\eta) = 0$, by taking $G = F_1$.

Convexity extends Eq. (15) to the global one-sided inequality

$$D(\gamma) \ge \kappa(\eta - \gamma)_+, \qquad 0 \le \gamma \le 1. \tag{22}$$

For $\gamma \in [\eta - t_0, \eta)$ this is the assumed bound. For $\gamma < \eta - t_0$, write $\eta - t_0 = s\gamma + (1-s)\eta$ with $s = t_0/(\eta - \gamma)$. Then $\kappa t_0 \le D(\eta - t_0) \le sD(\gamma)$. For $\gamma \ge \eta$, nonnegativity suffices.

Since $\widehat{D}(\widehat{\gamma}) \le \tau_n$, Lemma 1 gives

$$(\widehat{\pi} - \pi)_+ = (\eta - \widehat{\gamma})_+ \le \frac{1}{\kappa}\left\{\tau_n + \|\widehat{F} - F\|_\infty + \|\widehat{F}_0 - F_0\|_\infty + \frac{1}{n}\right\}. \tag{23}$$

The DKW–Massart inequality (Massart, 1990) states that an empirical CDF $\widehat{H}_k$ from k IID observations satisfies

$$\Pr\{\|\widehat{H}_k - H\|_\infty > x\} \le 2e^{-2kx^2}.$$

Applying it to the two samples and taking a union bound proves Eq. (16). For $\delta = 1/N$, use $\log(4N) \le 3\log N$, $\tau_n = 0.1\log\log(\max(n,10))/\sqrt{n} \le 0.1\sqrt{\log N/N}$, and $1/n \le \sqrt{\log N/N}$. The resulting coefficient is at most $0.1 + 2\sqrt{3/2} + 1 < 4$. □

We will also use an expectation bound and a feasibility bound. Integrating the DKW tail yields $\mathbb{E}\|\widehat{H}_k - H\|_\infty \le Ck^{-1/2}$ for a numerical constant C. Consequently, Eq. (23) implies

$$\mathbb{E}[(\widehat{\pi} - \pi)_+] \le \frac{C}{\kappa}(\tau_n + N^{-1/2}), \tag{24}$$

where C is numerical.

For feasibility at the true contamination proportion, use $G = F_1$ directly in Eq. (19):

$$\widehat{D}(\eta) \le \|\widehat{F} - F + \pi(F_0 - \widehat{F}_0)\|_{\widehat{F}} \le \|\widehat{F} - F\|_\infty + \pi\|\widehat{F}_0 - F_0\|_\infty.$$

If both empirical CDF errors are at most $\tau_n/2$, then $\widehat{D}(\eta) \le \tau_n$. On this event, the definition of $\widehat{\gamma}$ implies $\widehat{\gamma} \le \eta$ and hence $\widehat{\pi} \ge \pi$. Another union bound therefore gives, for every $\tau_n > 0$,

$$\Pr\{\widehat{\pi} < \pi\} \le \Pr\{\widehat{D}(\eta) > \tau_n\} \le 4\exp(-N\tau_n^2/2) \le 4\exp(-N\tau_n^2/8). \tag{25}$$

The last, looser form matches the remainder in Theorem 2. No independence between $\widehat{\pi}$ and the mixed observations is required.

Sufficiency of an anchor interval. Suppose $F_0(I) > 0$ and $F_1(I) = 0$ for an interval I. Continuity of F_0 allows ordered subintervals $A < B$ within I with $a = F(A) > 0$, $b = F(B) > 0$, and $\inf_{y\in B} F_0(y) - \sup_{x\in A} F_0(x) \ge \Delta > 0$. For $\gamma = \eta - t$ and any CDF G, the residual $U = F - (1-\gamma)F_0 - \gamma G$ satisfies $U(x) - U(y) \ge t\Delta$ for $x \in A, y \in B$. Jensen's inequality, followed by minimizing over the two conditional means, yields

$$\int U^2\, dF \ge \frac{ab}{a+b}t^2\Delta^2.$$

Taking the infimum over G proves Eq. (15) with $\kappa = \Delta\sqrt{ab/(a+b)}$.

A.3 Integrated KDE error with a reference-dependent bandwidth

Lemma 2 (Uniform integrated KDE bound)**.** *Let $X_1, \dots, X_k$ be IID with density g satisfying Assumption 3, and let $\widehat{g}_s = k^{-1} \sum_j K_s(\cdot - X_j)$. For $0 < h \le 1$,*

$$\mathbb{E} \sup_{s \in [uh, vh]} \|\widehat{g}_s - g\|_1 \le C \left(h^2 + \frac{1}{\sqrt{kh}} \right), \tag{26}$$

where C depends only on L, V, u, v.

Proof. Taylor's formula with integral remainder, $\int zK(z)\,dz = 0$, and $\int z^2 K(z)\,dz = 1$ give

$$\|\mathbb{E}\widehat{g}_s - g\|_1 = \|K_s * g - g\|_1 \le \frac{s^2}{2} \|g''\|_1 \le \frac{Ls^2}{2}.$$

For the centered process $Z_s = \widehat{g}_s - \mathbb{E}\widehat{g}_s$, weighted Cauchy–Schwarz and Jensen imply

$$\mathbb{E}\|Z_s\|_1 \le \left(\int \frac{dw}{1+w^2} \right)^{1/2} \left(\mathbb{E} \int (1+w^2) Z_s(w)^2 \, dw \right)^{1/2}.$$

Independence and a change of variables show

$$\begin{aligned} \mathbb{E} \int (1+w^2) Z_s(w)^2 \, dw &\le \frac{1}{k} \int (1+w^2) \mathbb{E} K_s(w - X_1)^2 \, dw \\ &= \frac{1}{ks} \mathbb{E} \int [1 + (X_1 + sz)^2] K(z)^2 \, dz \le \frac{C}{ks}, \end{aligned}$$

using $\mathbb{E}X_1^2 \le V$ and $s \le v$. Thus $\mathbb{E}\|Z_s\|_1 \le C(ks)^{-1/2}$.

To obtain uniformity in s, write $\partial_s K_s(x) = s^{-2} J(x/s)$, where $J(z) = -K(z) - zK'(z)$. The same calculation, using the finite Gaussian moments of J^2, gives

$$\mathbb{E} \int (1+w^2)(\partial_s Z_s(w))^2 \, dw \le \frac{C}{ks^3}, \qquad \mathbb{E}\|\partial_s Z_s\|_1 \le \frac{C}{\sqrt{k}\, s^{3/2}}.$$

The fundamental theorem of calculus and Tonelli's theorem yield

$$\begin{aligned} \mathbb{E} \sup_{s \in [uh, vh]} \|Z_s\|_1 &\le \mathbb{E}\|Z_{uh}\|_1 + \int_{uh}^{vh} \mathbb{E}\|\partial_s Z_s\|_1 \, ds \\ &\le \frac{C}{\sqrt{kh}}. \end{aligned}$$

Combining this with the uniform bias bound proves the lemma. □

In particular, the lemma controls $\widehat{f}_0$ even when σ is selected from the reference sample. It also controls $\widehat{f}$ and the leave-one-out KDE

$$\widehat{f}_{-i}(w) = \frac{1}{n-1} \sum_{j \ne i} K_\sigma(w - W_j).$$

Since $n - 1 \ge 2n/3$ and $m, n \ge N$, their expected L^1 errors are all at most $C\{h^2 + (Nh)^{-1/2}\}$. The restriction that σ not depend on the mixed sample is used below to preserve conditional independence of W_i and $\widehat{f}_{-i}$.

A.4 Average risk underestimation at the observed scores

For $a \in [0, 1]$, define

$$\ell(w) = \frac{\pi f_0(w)}{f(w)}, \qquad \widehat{\ell}_a(w) = \min \left\{ \frac{a\widehat{f}_0(w)}{\widehat{f}(w)}, 1 \right\}, \qquad \widehat{r}_a(w) = 1 - \widehat{\ell}_a(w).$$

Thus $r = 1 - \ell$ and $\widehat{r} = \widehat{r}_{\widehat{\pi}}$. We use the following elementary inequality: for $a \geq 0$, $b > 0$, $d > 0$, and $0 \leq c \leq d$,

$$d\,|\min\{a/b, 1\} - c/d| \leq |a - c| + |b - d|. \tag{27}$$

If $a \leq b$, then $|da/b - c| \leq |a - c| + (a/b)|d - b|$. If $a > b$, the left-hand side is $d - c$, and $d - c \leq |d - b| + (b - c)_+ \leq |d - b| + |a - c|$. Applied with $c = \pi f_0$ and $d = f$, this inequality avoids any positive lower bound on f.

Let $\widehat{\ell}_{\pi,-i}(w) = \min\{\pi \widehat{f}_0(w)/\widehat{f}_{-i}(w), 1\}$. Conditional on the reference and W_{-i}, the bandwidth is fixed and W_i still has density f. Integrating Eq. (27) therefore yields

$$\mathbb{E}|\widehat{\ell}_{\pi,-i}(W_i) - \ell(W_i)| \leq \pi \mathbb{E}\|\widehat{f}_0 - f_0\|_1 + \mathbb{E}\|\widehat{f}_{-i} - f\|_1. \tag{28}$$

The full-sample denominator satisfies

$$\widehat{f}(W_i) = \frac{n-1}{n}\widehat{f}_{-i}(W_i) + \frac{K(0)}{n\sigma}.$$

For $x \geq 0$ and $c \geq 1$, $\min\{cx, 1\} \leq \min\{x, 1\} + c - 1$. Consequently,

$$\widehat{\ell}_{\pi}(W_i) \leq \widehat{\ell}_{\pi,-i}(W_i) + \frac{1}{n-1}.$$

Monotonicity and clipping also give, for every w,

$$\widehat{\ell}_{\widehat{\pi}}(w) - \widehat{\ell}_{\pi}(w) \leq \frac{(\widehat{\pi} - \pi)_+}{\pi}.$$

Combining these inequalities and taking positive parts gives

$$\{r(W_i) - \widehat{r}(W_i)\}_+ \leq \frac{(\widehat{\pi} - \pi)_+}{\pi} + |\widehat{\ell}_{\pi,-i}(W_i) - \ell(W_i)| + \frac{1}{n-1}.$$

By Eqs. (24), (26), and (28),

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\{r(W_i) - \widehat{r}(W_i)\}_+\right] \leq C\left\{\frac{\tau_n + N^{-1/2}}{\kappa} + h^2 + \frac{1}{\sqrt{Nh}}\right\}. \tag{29}$$

Here C depends only on π, L, V, u, v; the term $1/(n-1)$ is absorbed because $h \leq 1$ and $N \geq 3$. The nonnegative self-kernel term in the mixture denominator is favorable for this one-sided comparison.

A.5 Irreducibility implies nonvanishing retention

Lemma 3 (Retention from a population low-risk region)**.** *Under Assumptions 1–3, define*

$$p = \Pr\{r(W) \leq \alpha/2\}. \tag{30}$$

Then $p > 0$, and

$$\Pr\{|R| < pn/4\} \leq C\left\{h^2 + \frac{1}{\sqrt{Nh}} + \exp(-N\tau_n^2/8)\right\}, \tag{31}$$

where C depends only on $\alpha, p, \pi, L, V, u, v$. The positivity of p is a consequence of irreducibility, not an additional assumption.

For the constant notation in Theorem 2, take $\varepsilon = \alpha/2$. Then $p = \Pr\{r(W) \leq \alpha - \varepsilon\}$, and dependence on α can equivalently be expressed as dependence on ε. Both quantities are fixed by the population and the target level; no extra assumption is imposed.

Proof. On $\{f_0 > 0\}$,

$$r(w) = \frac{(1-\pi)f_1(w)/f_0(w)}{\pi + (1-\pi)f_1(w)/f_0(w)}.$$

Irreducibility gives $\operatorname{ess\,inf}_{w \sim F_0} f_1(w)/f_0(w) = 0$. In particular, the set where $f_1/f_0 < \pi\alpha/\{(1-\pi)(2-\alpha)\}$ has positive F_0 probability and satisfies $r < \alpha/2$. Since $F \geq \pi F_0$ as measures, it also has positive F probability, proving $p > 0$.

Set $T = \sqrt{2(V+1)/p}$ and define

$$A = \{w : |w| \le T,\ r(w) \le \alpha/2\}, \qquad A_n = \{i : W_i \in A\}.$$

The moment bound gives $F(A) \ge p - V/T^2 \ge p/2$. This bounded set is only a witness to the existence of a feasible subset; the selected scores are not restricted to A. Introduce the average error on this set at the true clean proportion,

$$E_n = \frac{1}{n} \sum_{i \in A_n} |\widehat{r}_\pi(W_i) - r(W_i)|.$$

We claim

$$\mathbb{E}E_n \le C \left(h^2 + \frac{1}{\sqrt{Nh}} \right). \tag{32}$$

To prove this, condition on the reference and W_{-i}. As a function of the remaining query w, the full-sample mixture denominator evaluated at that query is

$$B_i(w) = \frac{n-1}{n} \widehat{f}_{-i}(w) + \frac{K(0)}{n\sigma}.$$

Equation (27), integrated over A with $a = \pi \widehat{f}_0(w)$, $b = B_i(w)$, $c = \pi f_0(w)$, and $d = f(w)$, gives

$$\begin{aligned} &\mathbb{E}\big[\mathbf{1}_{\{W_i \in A\}} |\widehat{r}_\pi(W_i) - r(W_i)| \mid C_{1:m}, W_{-i}, \mathcal{H}\big] \\ &\quad \le \pi \|\widehat{f}_0 - f_0\|_1 + \|\widehat{f}_{-i} - f\|_1 + \frac{1}{n} + \frac{2TK(0)}{n\sigma}. \end{aligned}$$

Here the last term is finite because $A \subseteq [-T, T]$. Taking expectations, using Lemma 2, $\sigma \ge uh$, and $(nh)^{-1} \le (Nh)^{-1} \le (Nh)^{-1/2}$ proves Eq. (32).

Since $|A_n|$ is binomial with success probability at least $p/2$, a multiplicative Chernoff bound gives

$$\Pr\{|A_n| < pn/4\} \le e^{-np/16}.$$

On the event $\{\widehat{\pi} \ge \pi,\ |A_n| \ge pn/4,\ E_n \le \alpha p/8\}$, monotonicity in the clean proportion implies

$$\frac{1}{|A_n|} \sum_{i \in A_n} \widehat{r}(W_i) \le \frac{1}{|A_n|} \sum_{i \in A_n} \widehat{r}_\pi(W_i) \le \frac{\alpha}{2} + \frac{n}{|A_n|} E_n \le \alpha.$$

Thus A_n is a feasible subset for the average estimated-risk constraint. Among all subsets of a given size, the prefix with the smallest estimated risks has the smallest sum. The maximal feasible prefix therefore satisfies $|R| \ge |A_n|$ on this event. By Eq. (25) and Markov's inequality,

$$\Pr\{|R| < pn/4\} \le 4e^{-N\tau_n^2/8} + e^{-np/16} + \frac{8}{\alpha p} \mathbb{E}E_n.$$

Finally, $e^{-np/16} \le C_p n^{-1/2} \le C_p (Nh)^{-1/2}$ because $N \le n$ and $h \le 1$. Substitution of Eq. (32) proves the lemma. □

A.6 Proof of the finite-sample FDR bound

Proof of Theorem 2. The selection rule is measurable with respect to the observed scores. Under Assumption 1, the conditional probability that observation i is contaminated is $r(W_i)$, even after conditioning on all other scores and the reference sample. Hence

$$\mathbb{E}[\mathrm{FDP}(R) \mid W_{1:n}, C_{1:m}, \mathcal{H}] = \frac{\sum_{i \in R} r(W_i)}{|R| \vee 1}. \tag{33}$$

The prefix rule ensures $|R|^{-1} \sum_{i \in R} \widehat{r}(W_i) \le \alpha$ whenever R is nonempty. On $\{|R| \ge pn/4\}$, the right-hand side of Eq. (33) is consequently at most

$$\alpha + \frac{4}{p} \frac{1}{n} \sum_{i=1}^{n} \{r(W_i) - \widehat{r}(W_i)\}_+.$$

For an empty selection FDP is zero; on $\{0 < |R| < pn/4\}$ it is at most one. Taking expectations therefore gives

$$\mathrm{FDR}(R) \le \alpha + \frac{4}{p}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\{r(W_i) - \widehat{r}(W_i)\}_+\right] + \Pr\{0 < |R| < pn/4\}. \tag{34}$$

Combining Eq. (29) with Lemma 3 yields

$$\mathrm{FDR}(R) \le \alpha + C\left\{\frac{\tau_n + N^{-1/2}}{\kappa} + h^2 + \frac{1}{\sqrt{Nh}} + \exp(-N\tau_n^2/8)\right\},$$

which is Eq. (17). The constant depends only on $\pi, p, \varepsilon, L, V, u, v$, where $\varepsilon = \alpha/2$ and the positive number p in Eq. (30) is determined by the fixed population and target level. In particular, the theorem is pointwise in that population; no uniform lower bound on p over a class of populations is asserted. There is no dependence on m, n, h, τ_n, κ beyond the displayed factors.

For a fixed population, $h \to 0$, $Nh \to \infty$, $\tau_n \to 0$, and $N\tau_n^2 \to \infty$ make all remainder terms vanish, proving Eq. (18). □

With $h = N^{-1/5}$, both KDE terms equal $N^{-2/5}$. For the original tolerance in Eq. (35), $m \asymp n$ gives $N\tau_n^2 \asymp (\log\log n)^2 \to \infty$ and $\tau_n \to 0$. This proves the stated asymptotic interpretation while retaining the feasibility remainder in the finite-sample bound.

B Experimental Details

B.1 Datasets and Models

We evaluate Pythia-2.8B, OPT-6.7B, and LLaMA-13B on WikiMIA, arXivTection, and BBC Real Time (news), constructing the sample pools using the temporal splits and membership labels of the existing benchmarks (Shi et al., 2024; Zhang et al., 2026). The datasets are briefly described below.

- **WikiMIA.** This benchmark (Shi et al., 2024) contains Wikipedia event texts and assigns membership labels using temporal information. Events before 2017 are treated as contaminated data, and events after 2023 as clean data.
- **arXivTection.** This benchmark (Duarte et al., 2024) comprises 50 research papers from arXiv and evaluates training data detection in the scientific domain. Papers published before 2022 are labeled contaminated, whereas papers published in 2023 are labeled clean.
- **BBC Real Time.** This dataset (Li et al., 2024) contains BBC news articles published between January 2017 and August 2024. Following Shi et al. (2024), articles from 2017 are used as contaminated data and articles from 2024 as clean data.

These temporal splits define the benchmark sample pools, while designated contamination exposure in the controlled experiments is introduced through subsequent fine-tuning.

We introduce controlled contamination through LoRA fine-tuning (Hu et al., 2022) on designated member samples, followed by continued training on independent clean data to construct the fsd scores used in the experiments. The clean reference set $\mathcal{D}_{\mathrm{cal}}$ is used for distribution estimation; the auxiliary clean set A_{clean} is used for clean continued training and fitting the combined-score transformations; and the candidate evaluation set $\mathcal{D}_{\mathrm{test}}$ is used for selection and evaluation. Sample counts are reported in Table 3. The selection methods do not use candidate membership labels or the true clean proportion.

Both training stages use LoRA rank $r_{\mathrm{LoRA}} = 16$, scaling parameter $\alpha_{\mathrm{LoRA}} = 32$, a batch size of 8, and a cosine learning-rate schedule. Contamination fine-tuning and clean continued training use dropout rates of 0.10 and 0.05, 10 and 3 epochs, and learning rates of 5×10^{-5} and 10^{-3}, respectively.

B.2 Scoring Functions and Baselines

We use the following nine scoring configurations:

Table 3: Data partitions and sample counts for the main experiments. The candidate evaluation set contains clean samples and samples used in the designated contamination training.

Dataset	Contamination training samples	$\lvert\mathcal{D}_{\mathrm{cal}}\rvert$	$\lvert A_{\mathrm{clean}}\rvert$	Clean candidates	Contaminated candidates
WikiMIA	861	394	121	274	304
arXivTection	762	393	123	270	272
news	3,276	1,673	499	1,174	1,144

- **PPL.** The model's perplexity on the input text is used as the training data detection score (Li, 2023).
- **Lowercase.** We take the negative ratio of the mean negative log-likelihood of the lowercased text to that of the original text to align the score direction (Carlini et al., 2021).
- **Zlib.** The score is the ratio of the mean negative log-likelihood of the original text to its zlib-compressed byte length (Carlini et al., 2021).
- **Grad.** The score is the L_2 norm of the gradient of the mean prediction loss with respect to the model parameters (Hu et al., 2025).
- **Min-K 20% and Min-K 30%.** We select the 20% and 30% of tokens with the lowest probabilities, respectively, and use the negative of their mean log probability (Shi et al., 2024).
- **PAC.** Following Polarized Augment Calibration (Ye et al., 2024), we compare the probability discrepancy of the original text with that of five word-swap perturbations. The discrepancy is defined as the difference between the mean log probabilities of the top 5% and bottom 30% of tokens.
- **Combined scores.** In addition to individual scores, we construct MEAN_Z and First_PCA using standardized averaging and first principal-component projection, respectively. Specifically, we standardize six scores—PPL, Lowercase, Zlib, Grad, Min-K 20%, and Min-K 30%—and take their mean and first principal-component projection (Jolliffe & Cadima, 2016), respectively. The standardization parameters and principal-component direction are fitted only on the independent auxiliary clean set and remain fixed during subsequent selection.

Texts are scored using at most 512 tokens, and score directions are fixed before selection. The compared methods are Raw, fsd, and CleanRef-LFDR, as defined in the main text. The fsd scores are constructed from score differences before and after clean continued training (Zhang et al., 2025a). In the main comparison, fsd and CleanRef-LFDR use the same fsd scores and differ in their selection rules. PU-BH strictly follows the Stage I procedure in Section 4.2 of He et al. (2026) and the accompanying source implementation, using a random forest (Breiman, 2001), two-component Gaussian-mixture calibration, and BH selection (Benjamini & Hochberg, 1995). The random forest contains 100 trees. All methods use the same clean reference set, candidate evaluation set, and target risk level.

B.3 Implementation and Evaluation Protocol

Unless the target level is varied, we set $\alpha = 0.15$.

For the main comparison, repeated evaluations keep the trained models and scores fixed and use paired candidate samples across methods; the detailed repetition protocol is provided in the released code.

CleanRef-LFDR applies shared standardization based on the clean reference scores, estimates the contamination proportion $\widehat{\gamma}$ by continuous bisection, and obtains the clean-proportion estimate as $\widehat{\pi} = 1 - \widehat{\gamma}$. The search stops at an interval width of 10^{-8}. The projection tolerance is

$$\tau_n = \frac{0.1 \log\log(\max(n, 10))}{\sqrt{n}}, \tag{35}$$

where n is the number of candidates. Both Gaussian KDEs use 1.5 times the NRD0 bandwidth of the reference scores, with an NRD0 numerical floor of 10^{-3}. The implementation clips the clean-proportion estimate used in risk calculation to $[10^{-6}, 1-10^{-6}]$, computes the density ratio in the log domain, and applies the average-risk rule with a comparison tolerance of 10^{-12}. The theory analyzes the corresponding statistical estimator; these are the numerical settings used in the experiments.

The computing environment for model training and scoring is equipped with eight NVIDIA A800-SXM4 GPUs, each with 80 GB of memory, and 105 vCPUs. Statistical selection runs on CPUs using NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), and scikit-learn (Pedregosa et al., 2011). The implementation uses PyTorch (Paszke et al., 2019) and Hugging Face Transformers (Wolf et al., 2020), with AdamW (Loshchilov & Hutter, 2019) for LoRA fine-tuning.

Stability across repeated evaluations. Over 100 repetitions, Monte Carlo standard errors of empirical FDR for CleanRef-LFDR range from 0.0016 to 0.0050. In every configuration where both CleanRef-LFDR and fsd meet the target, the per-configuration 95% paired Monte Carlo interval for the power difference lies above zero. These results support precise mean estimates and a consistent advantage in clean-sample retention under the evaluated conditions.

C Additional Sensitivity Results

We report all nine scoring configurations across the three models and three datasets. Unless explicitly varied, $\alpha = 0.15$. The matrices retain each score–model–dataset–parameter configuration separately, without averaging across scores. Black outlines mark results whose empirical FDR exceeds the corresponding target. Power is shown on a separate color scale.

C.1 Target Risk Level

We vary only the target risk level while keeping the other conditions fixed to examine changes in selection performance. Figure 4 includes both baselines and CleanRef-LFDR.

C.2 Clean-Proportion Perturbation

To isolate the effect of proportion estimation, we fix both density estimates and replace the estimated clean proportion by

$$\widetilde{\pi}(\delta_\pi) = \mathrm{clip}_{[10^{-6},\,1-10^{-6}]}\big(\widehat{\pi} + \delta_\pi\big), \qquad \delta_\pi \in \{-0.10, -0.05, 0, 0.05, 0.10\}.$$

The unperturbed point uses the same fit as the main comparison. Because the baselines do not use this estimate, this intervention reports CleanRef-LFDR only. Figure 5 shows First_PCA on LLaMA-13B, and Figure 6 reports all scores.

The original proportion estimate already supports near-perfect clean-sample recovery in the displayed configurations. Reducing the estimate makes selection more conservative, whereas increasing it raises empirical FDR with little additional gain in power. These results suggest that, once clean samples are largely recovered, further increasing the estimated clean proportion offers limited benefit. Accurate contamination-risk estimation remains important for making effective use of the risk budget. Appendix E.2 further separates proportion and density effects using known population quantities.

C.3 Clean-Reference Size

To examine selection with a limited clean reference, we keep the candidate evaluation set fixed and use 25%, 50%, 75%, and 100% of the original reference size. Complete results are shown in Figure 7.

Even with one quarter of the reference data, CleanRef-LFDR keeps empirical FDR below the target and retains more clean samples than both baselines in most configurations. As the reference size increases, power improves in most configurations and approaches one in several settings. These results show that a limited clean reference can already support effective selection, while additional

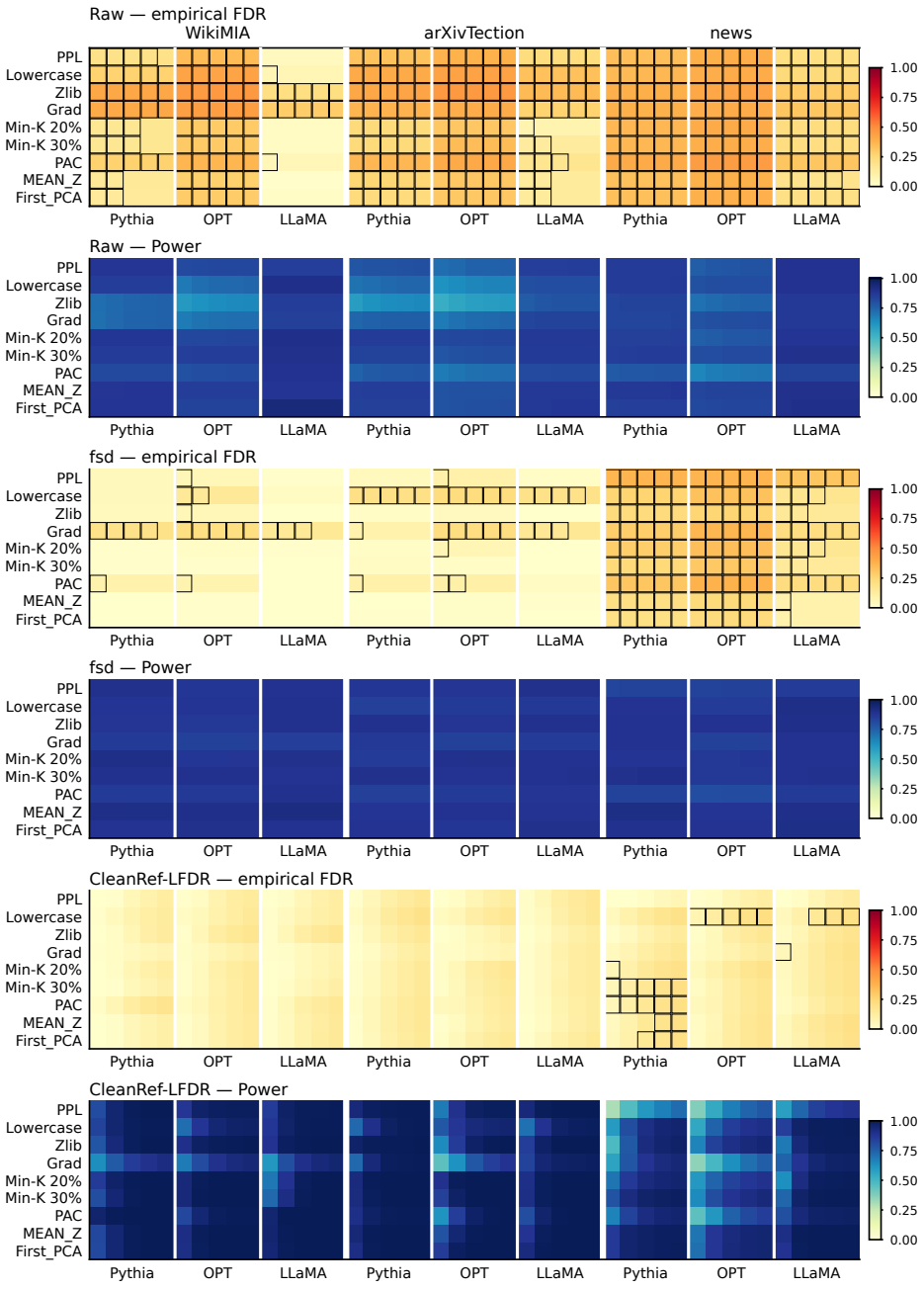

Figure 4: Complete target-level analysis. Each row corresponds to a score, and columns are grouped by dataset and model. Within each model, the five cells correspond to $\alpha = 0.05, 0.10, 0.15, 0.18, 0.20$. The matrices report empirical FDR and Power for each method.

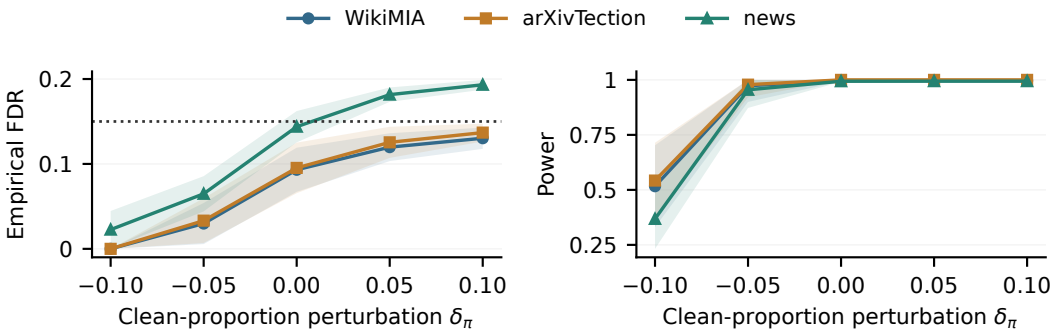

Figure 5: Clean-proportion perturbation for First_PCA on LLaMA-13B at $\alpha = 0.15$. Curves correspond to different datasets, and $\delta_\pi = 0$ denotes the original estimate. Shaded bands indicate one standard deviation. The dotted line denotes the risk target.

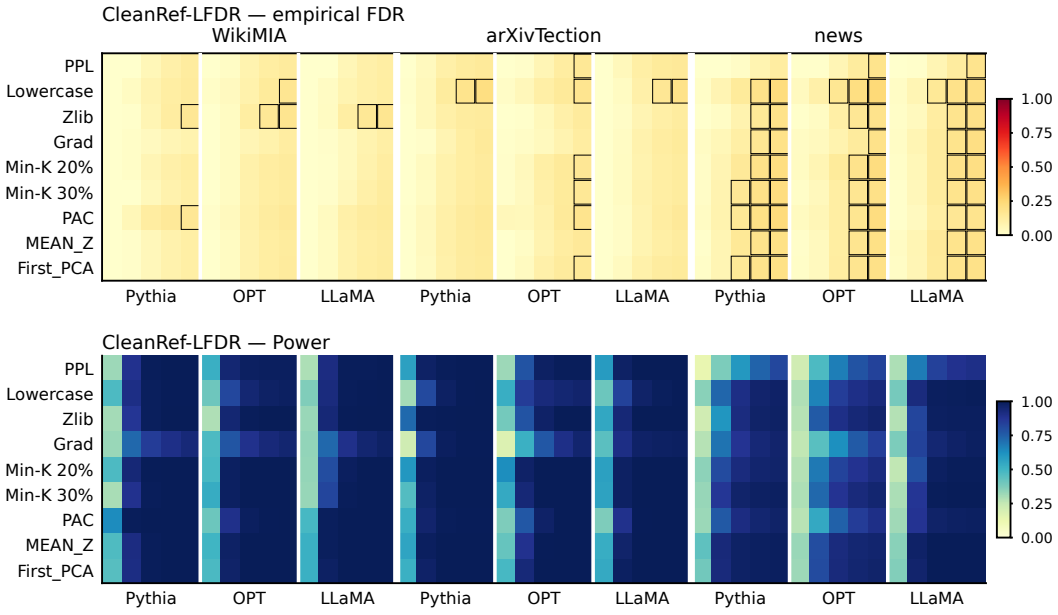

Figure 6: Complete clean-proportion intervention for CleanRef-LFDR at $\alpha = 0.15$. Within each model–dataset block, columns correspond to $\delta_\pi = -0.10, -0.05, 0, 0.05, 0.10$.

reference information can further improve clean-sample recovery. The three datasets displayed in Figure 3 show the same pattern: CleanRef-LFDR maintains low empirical FDR and high power across reference sizes.

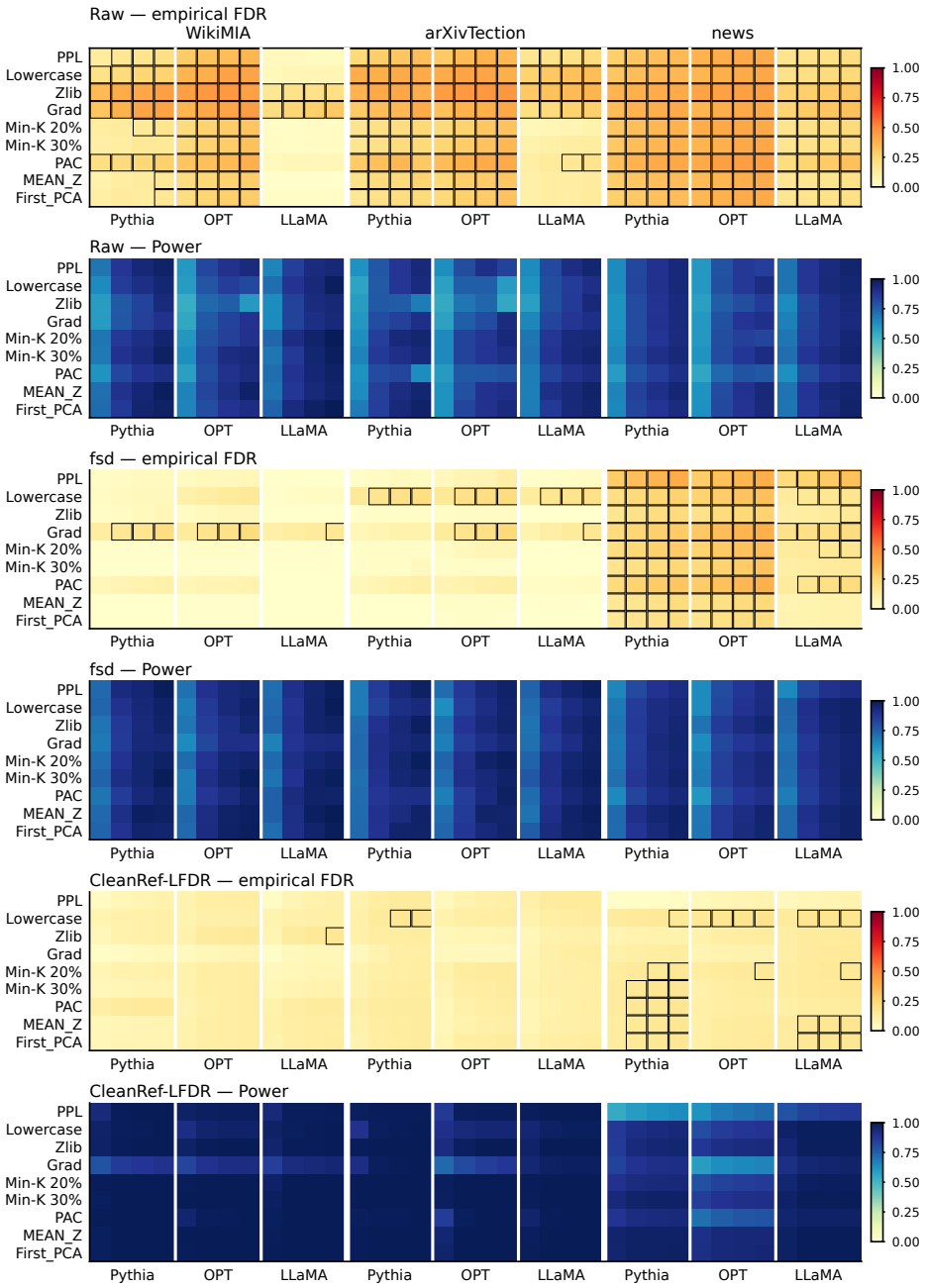

Figure 7: Effect of clean-reference size on empirical FDR and Power at $\alpha = 0.15$. Within each model–dataset block, columns correspond to 25%, 50%, 75%, and 100% of the original reference size. Each scoring configuration is reported separately.

C.4 Reference-Domain Composition

Effect of reference-domain composition. We use Min-K 30% on LLaMA-13B/news to examine selection under changes in reference sources. We first replace the original in-domain reference with

newly collected BBC data, then progressively replace up to 50% of these samples with Wikipedia data while keeping the total reference size fixed.

Figure 8(a) shows that, after replacing the in-domain reference, CleanRef-LFDR maintains empirical FDR below the target and higher power than both baselines. As the Wikipedia replacement proportion increases, empirical FDR decreases further. Although power declines, it remains above both baselines at every displayed proportion [Figure 8(b–c)]. These results show that CleanRef-LFDR continues to limit residual contamination and retains its advantage in clean-sample recovery under the evaluated changes in reference sources. Results for other models and scores are reported below.

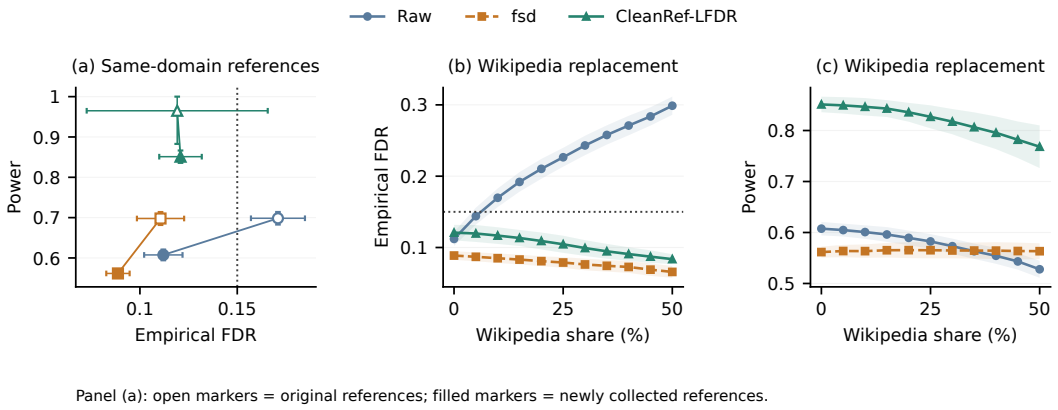

Figure 8: Effects of in-domain reference replacement and out-of-domain reference inclusion, using Min-K 30% on LLaMA-13B/news at $\alpha = 0.15$. (a) Comparison of equally sized original and new BBC reference sets, shown by open and filled markers, respectively. (b–c) Empirical FDR and Power as the Wikipedia replacement proportion increases from 0% to 50%. Error bars and shaded bands indicate one standard deviation. The dotted line denotes the risk target.

We further examine how changes in reference sources affect selection. Figure 9 reports results for MEAN_Z and First_PCA across the three models on news. Wikipedia data replace 0–50% of the in-domain clean reference.

The new BBC reference consists of text excerpts extracted from 2024 news snapshots, with publication in 2024 verified using the original publication dates on the official BBC pages. The new Wikipedia reference is drawn from the nonmember texts in WikiMIA-24; page creation in 2024 is verified through the official Wikipedia API. Source links and the corresponding temporal evidence are retained for both reference sets.

In every displayed configuration in Figure 9, CleanRef-LFDR keeps empirical FDR below the target and achieves higher power than both baselines at all replacement proportions. Although power varies with reference composition to different degrees across models, the advantages in risk control and clean-sample retention persist. These results show that CleanRef-LFDR can continue to limit residual contamination and retain more clean evaluation data under the evaluated partial inclusion of out-of-domain references.

Table 4 further examines complete replacement with Wikipedia references. For First_PCA on LLaMA-13B/news, empirical FDR stays below the target throughout Wikipedia replacement. Even with complete replacement, CleanRef-LFDR retains more clean samples than both baselines.

C.5 Effect of Test-Set Contamination Proportion

To examine how CleanRef-LFDR adapts to different contamination levels, we fix the clean reference set and total test size on arXivTection and vary the test contamination proportion over 10%, 25%, 50%, 75%, and 90%. The experiments cover Pythia-2.8B, OPT-6.7B, and LLaMA-13B, with all nine scoring configurations and a target level of $\alpha = 0.15$. The three figures report empirical FDR and Power for each model.

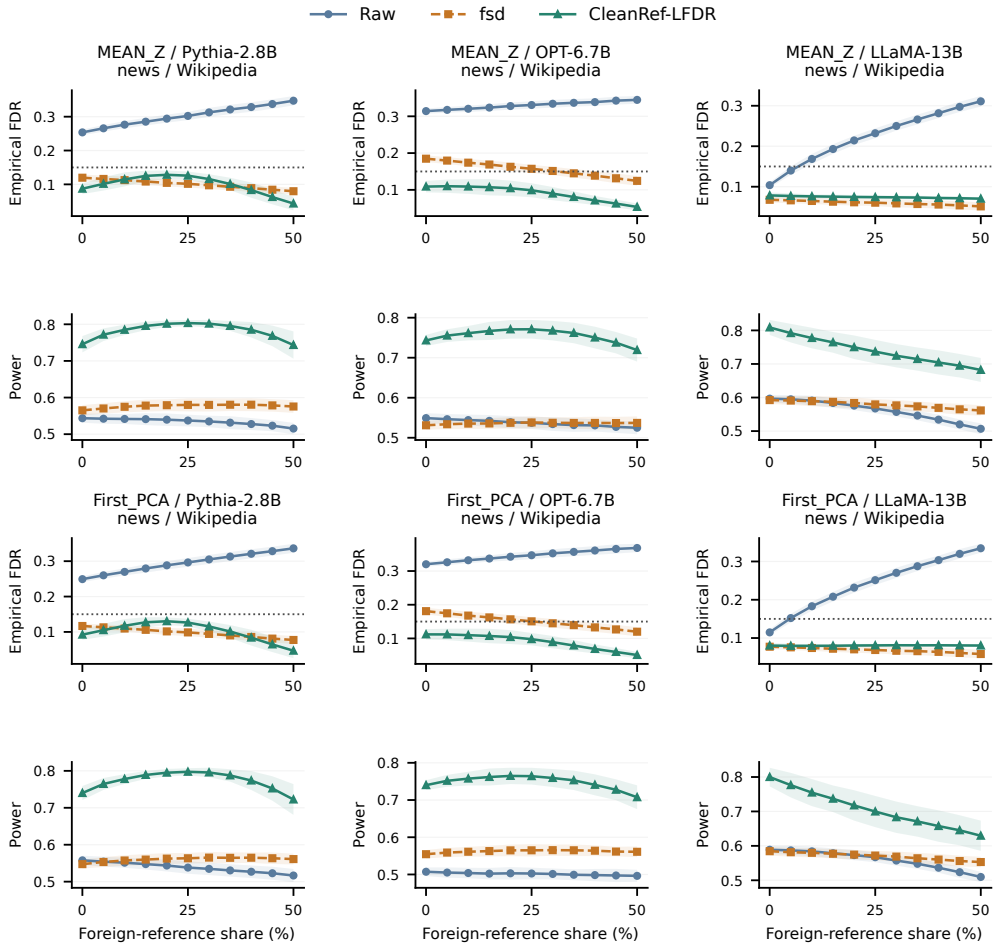

Figure 9: Effect of reference-domain composition on empirical FDR and Power at $\alpha = 0.15$. Results cover two combined scores and three models on news. The horizontal axis gives the Wikipedia replacement proportion, and shaded bands indicate one standard deviation.

Table 4: Selection results for First_PCA on LLaMA-13B/news under different reference compositions at $\alpha = 0.15$. Each cell reports empirical FDR / Power. Bold follows the criterion in Table 1.

Reference composition	Raw	fsd	CleanRef-LFDR
BBC only	0.115 / 0.589	0.077 / 0.584	**0.080 / 0.800**
25% Wikipedia	0.251 / 0.567	0.069 / 0.572	**0.081 / 0.700**
50% Wikipedia	0.335 / 0.510	0.058 / 0.553	**0.080 / 0.629**
100% Wikipedia	0.555 / 0.261	0.030 / 0.507	**0.080 / 0.512**

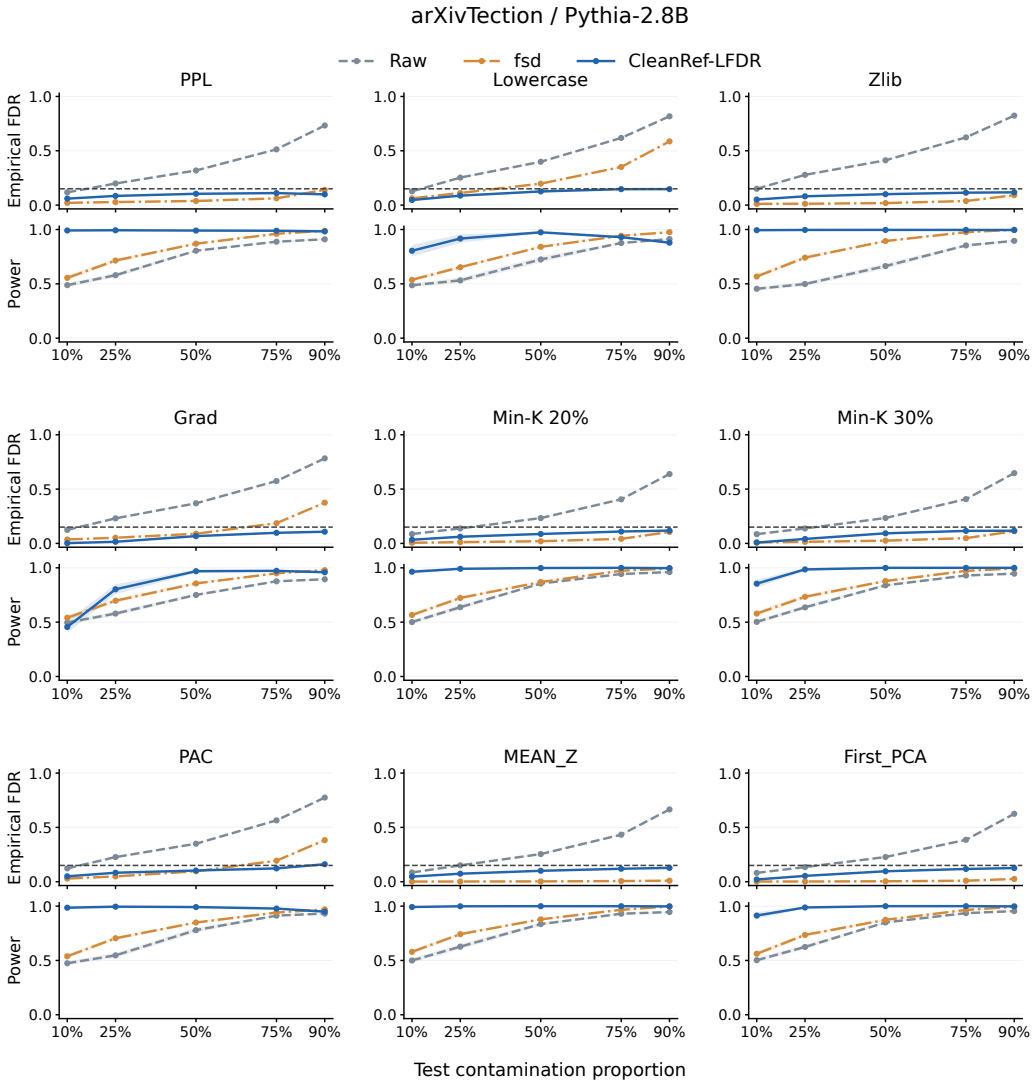

Figure 10: Effect of test contamination proportion on selection performance on arXivTection at $\alpha = 0.15$. Results are shown for Pythia-2.8B with all nine scoring configurations. For each score, the upper and lower panels report empirical FDR and Power. The horizontal axis gives the proportion of contaminated test samples. Curves show average results, and shaded bands indicate 95% Monte Carlo intervals for the means. Horizontal dashed lines in the FDR panels denote the target level. The clean reference set and total test size remain fixed.

Stable risk control. At input contamination proportions of 25% and 50%, CleanRef-LFDR keeps empirical FDR below the target for every model and scoring configuration. Most configurations continue to meet the target at higher contamination levels. MEAN_Z and First_PCA maintain control across all three models and all five proportions, demonstrating robustness to changes in test-set composition within these settings. Since 10% input contamination is already below the risk target, this condition primarily examines clean-sample retention under low contamination.

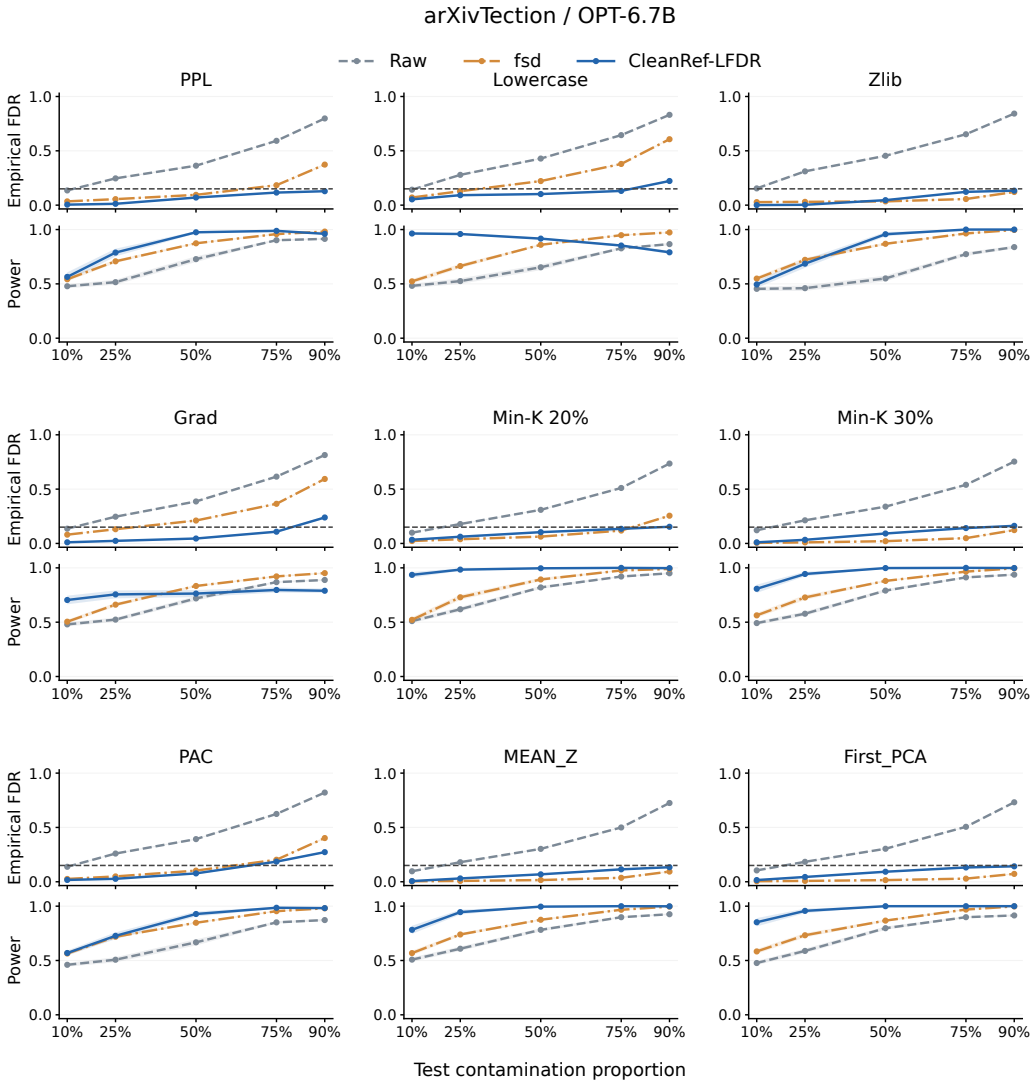

Figure 11: Effect of test contamination proportion on selection performance on arXivTection at $\alpha = 0.15$. Results are shown for OPT-6.7B with all nine scoring configurations. For each score, the upper and lower panels report empirical FDR and Power. The horizontal axis gives the proportion of contaminated test samples. Curves show average results, and shaded bands indicate 95% Monte Carlo intervals for the means. Horizontal dashed lines in the FDR panels denote the target level. The clean reference set and total test size remain fixed.

High clean-sample recovery. While maintaining low residual contamination, CleanRef-LFDR achieves higher power than both baselines in most configurations and near-perfect power in several settings. In particular, at moderate to high contamination levels, the method still recovers most clean samples even as their share of the test set decreases. With MEAN_Z and First_PCA, this power advantage persists across all three models and all five contamination levels.

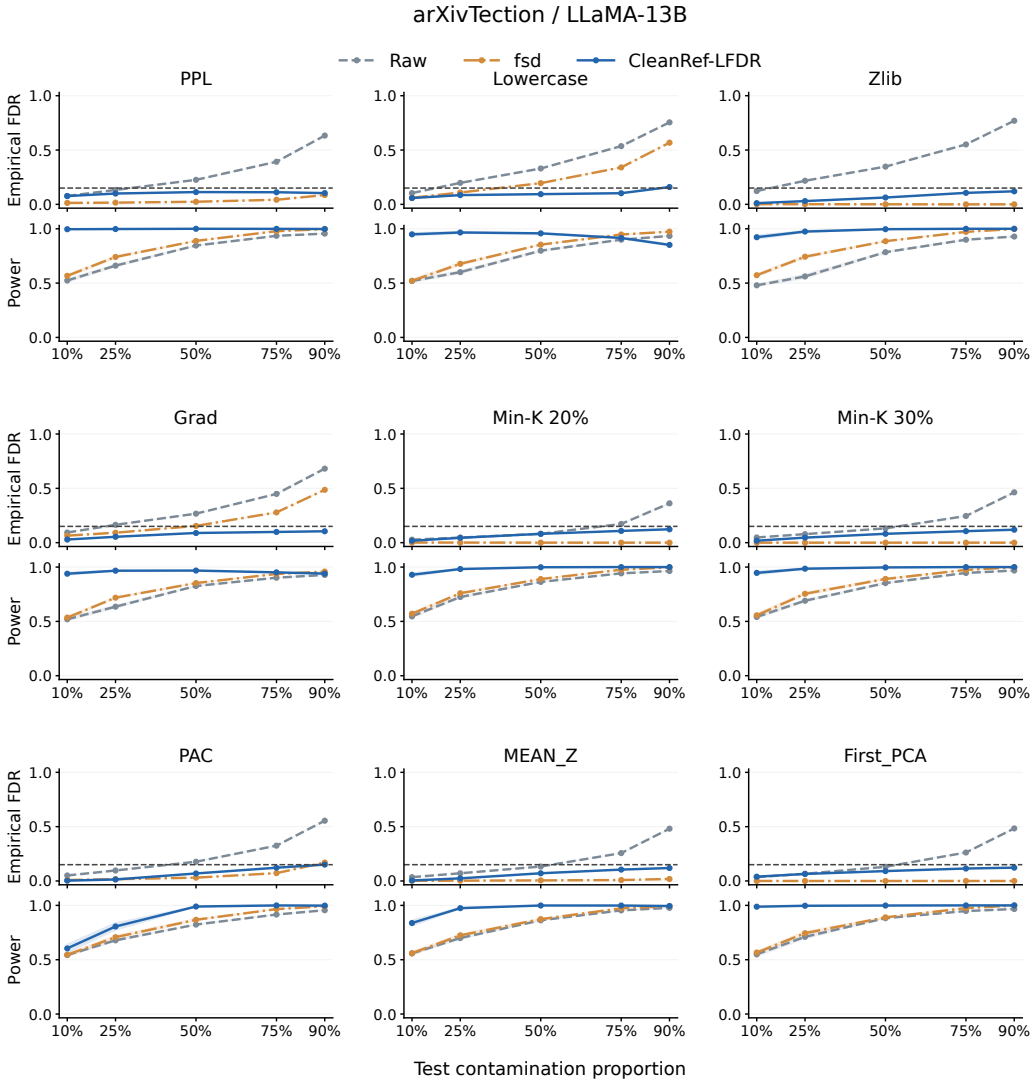

Figure 12: Effect of test contamination proportion on selection performance on arXivTection at $\alpha = 0.15$. Results are shown for LLaMA-13B with all nine scoring configurations. For each score, the upper and lower panels report empirical FDR and Power. The horizontal axis gives the proportion of contaminated test samples. Curves show average results, and shaded bands indicate 95% Monte Carlo intervals for the means. Horizontal dashed lines in the FDR panels denote the target level. The clean reference set and total test size remain fixed.

A better balance between risk and power. As test contamination increases, residual contamination under Raw rises substantially, making it difficult to consistently meet the target risk level. Although fsd maintains low empirical FDR in some configurations, it often recovers fewer clean samples. In comparison, CleanRef-LFDR combines low empirical FDR with high power in most settings. This suggests that selection based on contamination risk makes more effective use of clean reference information, retaining more data for reliable evaluation as test contamination changes.

D Additional Results on MIMIR

Members and nonmembers in MIMIR originate from the same corpora (Gao et al., 2020; Duan et al., 2024), making training data detection challenging when membership signals are weak (Jagielski et al., 2023; Bordt et al., 2025). Previous studies report near-random discrimination in several settings alongside relative performance improvements (Duan et al., 2024; Xie et al., 2024). We use this benchmark to further examine the ability of CleanRef-LFDR to reduce residual contamination while retaining clean samples.

D.1 Experimental Setup

We evaluate Pythia-2.8B-deduped on seven MIMIR domains using ReCaLL, First_PCA, and Grad. For each domain, we introduce controlled contamination by LoRA fine-tuning (Hu et al., 2022) on 500 member samples. These samples are combined with 300 clean samples to form the candidate evaluation set, with a separate clean reference set of 500 samples and an auxiliary set of 50 samples.

LoRA uses rank $r_{\mathrm{LoRA}} = 16$, scaling parameter $\alpha_{\mathrm{LoRA}} = 32$, and dropout 0.05. Training runs for 3 epochs with a batch size of 8, a learning rate of 10^{-3}, and a cosine learning-rate schedule.

ReCaLL (Xie et al., 2024) uses in-domain samples from the auxiliary set as prefixes. The six-feature transformation for First_PCA is fitted on the auxiliary set, and Grad follows the definition in Appendix B.2. This experiment compares PU-BH and CleanRef-LFDR using the same target-model scores, clean reference set, and candidate evaluation set.

D.2 Results and Analysis

Table 5 reports the complete results at the default target level $\alpha = 0.15$, and Table 6 provides comparisons at other target levels.

Score information affects risk control. CleanRef-LFDR achieves lower empirical FDR than PU-BH across all seven domains and all three scoring configurations, while attainment of the target risk level varies with the scoring configuration. On DM Mathematics, ReCaLL and First_PCA do not meet the target, whereas Grad keeps empirical FDR below the target while retaining a substantial proportion of clean samples. These results show that CleanRef-LFDR can use informative scores to improve the trade-off between risk and power, while its performance remains dependent on the available score information.

Risk level and power. Increasing the target level allows CleanRef-LFDR to retain more clean samples, whereas a lower target produces stricter selection. Table 6 illustrates this relationship between risk and power, showing that local-risk selection responds to different risk budgets.

Role of proportion estimation. To examine how clean-proportion estimation affects selection, we replace the estimated proportion with the construction value $\pi = 0.375$, holding all other conditions fixed. In settings where the clean proportion is overestimated, this substitution typically lowers empirical FDR and makes selection more conservative. These results show that proportion estimation directly affects contamination-risk assessment. Accurately characterizing the clean share of the mixture is therefore important for balancing risk control and power. Some configurations remain above the target after true-proportion substitution, indicating that proportion estimation alone does not explain all remaining error. The true proportion is used only for this diagnostic.

Table 5: Selection results and true-proportion diagnostics on MIMIR at $\alpha = 0.15$. Each cell reports empirical FDR / Power. $\Delta\pi$ is the mean of $\widehat{\pi} - \pi$, where $\pi = 0.375$. The true-proportion diagnostic replaces only the proportion term in risk estimation.

			Selection methods (raw scores)		Diagnostic
Score	Domain	$\Delta\pi$	PU-BH	CleanRef-LFDR	True π
ReCaLL	DM Mathematics	$+0.386$	0.626 / 0.527	0.565 / 0.592	0.435 / 0.005
ReCaLL	GitHub	$+0.033$	0.412 / 0.841	0.191 / 0.756	0.113 / 0.625
ReCaLL	Pile-CC	$+0.012$	0.450 / 0.823	0.073 / 0.755	0.052 / 0.728
ReCaLL	PubMed Central	-0.004	0.411 / 0.791	0.117 / 0.793	0.119 / 0.799
ReCaLL	arXiv	$+0.018$	0.284 / 0.864	0.137 / 0.928	0.112 / 0.918
ReCaLL	HackerNews	$+0.059$	0.462 / 0.805	0.244 / 0.809	0.164 / 0.685
ReCaLL	Wikipedia	-0.004	0.534 / 0.737	0.128 / 0.640	0.126 / 0.645
First_PCA	DM Mathematics	$+0.425$	0.609 / 0.569	0.584 / 0.929	0.010 / 0.000
First_PCA	GitHub	$+0.103$	0.463 / 0.819	0.292 / 0.669	0.209 / 0.197
First_PCA	Pile-CC	$+0.075$	0.527 / 0.762	0.219 / 0.649	0.140 / 0.341
First_PCA	PubMed Central	$+0.008$	0.478 / 0.742	0.170 / 0.603	0.160 / 0.578
First_PCA	arXiv	$+0.031$	0.483 / 0.759	0.180 / 0.610	0.127 / 0.525
First_PCA	HackerNews	$+0.067$	0.451 / 0.841	0.197 / 0.733	0.110 / 0.411
First_PCA	Wikipedia	$+0.080$	0.550 / 0.730	0.268 / 0.658	0.148 / 0.340
Grad	DM Mathematics	$+0.068$	0.523 / 0.765	0.134 / 0.699	0.046 / 0.366
Grad	GitHub	$+0.023$	0.399 / 0.809	0.163 / 0.712	0.144 / 0.625
Grad	Pile-CC	$+0.087$	0.423 / 0.816	0.256 / 0.850	0.178 / 0.424
Grad	PubMed Central	$+0.006$	0.419 / 0.790	0.092 / 0.695	0.082 / 0.686
Grad	arXiv	$+0.010$	0.297 / 0.864	0.094 / 0.885	0.082 / 0.876
Grad	HackerNews	$+0.075$	0.469 / 0.789	0.218 / 0.784	0.148 / 0.542
Grad	Wikipedia	$+0.033$	0.298 / 0.843	0.138 / 0.916	0.096 / 0.850

Table 6: MIMIR selection results at different target levels. Each cell reports empirical FDR / Power, with the three scoring configurations shown separately.

		$\alpha = 0.05$		$\alpha = 0.10$		$\alpha = 0.20$	
Score	Domain	PU-BH	CleanRef-LFDR	PU-BH	CleanRef-LFDR	PU-BH	CleanRef-LFDR
ReCaLL	DM Mathematics	0.626/0.479	0.514/0.318	0.626/0.508	0.542/0.445	0.628/0.540	0.591/0.762
ReCaLL	GitHub	0.412/0.841	0.070/0.503	0.412/0.841	0.135/0.655	0.412/0.841	0.241/0.835
ReCaLL	Pile-CC	0.449/0.820	0.020/0.529	0.450/0.823	0.039/0.655	0.450/0.823	0.128/0.829
ReCaLL	PubMed Central	0.411/0.791	0.040/0.620	0.411/0.791	0.077/0.729	0.411/0.791	0.156/0.844
ReCaLL	arXiv	0.284/0.864	0.035/0.825	0.284/0.864	0.082/0.895	0.284/0.864	0.191/0.950
ReCaLL	HackerNews	0.462/0.801	0.157/0.668	0.462/0.805	0.200/0.755	0.462/0.805	0.288/0.847
ReCaLL	Wikipedia	0.531/0.724	0.039/0.519	0.534/0.736	0.088/0.585	0.534/0.739	0.173/0.683
First_PCA	DM Mathematics	0.607/0.530	0.495/0.615	0.609/0.554	0.546/0.799	0.610/0.583	0.614/0.989
First_PCA	GitHub	0.462/0.817	0.235/0.408	0.462/0.818	0.264/0.548	0.463/0.819	0.324/0.773
First_PCA	Pile-CC	0.524/0.744	0.155/0.446	0.527/0.758	0.188/0.564	0.528/0.766	0.258/0.716
First_PCA	PubMed Central	0.478/0.740	0.108/0.410	0.478/0.742	0.140/0.514	0.478/0.742	0.208/0.673
First_PCA	arXiv	0.483/0.759	0.094/0.423	0.483/0.759	0.126/0.524	0.483/0.759	0.234/0.689
First_PCA	HackerNews	0.451/0.839	0.123/0.463	0.451/0.841	0.157/0.614	0.451/0.841	0.240/0.826
First_PCA	Wikipedia	0.549/0.715	0.171/0.512	0.549/0.728	0.217/0.598	0.550/0.734	0.316/0.708
Grad	DM Mathematics	0.522/0.756	0.058/0.448	0.523/0.762	0.088/0.606	0.523/0.766	0.188/0.757
Grad	GitHub	0.399/0.809	0.113/0.475	0.399/0.809	0.139/0.608	0.399/0.809	0.190/0.797
Grad	Pile-CC	0.423/0.816	0.194/0.650	0.423/0.816	0.218/0.782	0.423/0.816	0.293/0.895
Grad	PubMed Central	0.419/0.790	0.032/0.514	0.419/0.790	0.057/0.615	0.419/0.790	0.142/0.754
Grad	arXiv	0.297/0.864	0.036/0.689	0.297/0.864	0.057/0.816	0.297/0.864	0.143/0.924
Grad	HackerNews	0.469/0.787	0.151/0.560	0.469/0.789	0.180/0.687	0.469/0.789	0.266/0.846
Grad	Wikipedia	0.298/0.843	0.084/0.699	0.298/0.843	0.100/0.841	0.298/0.843	0.190/0.951

E Additional Statistical Analyses

E.1 Effect of the Bandwidth

To examine how density smoothing affects selection, we set the shared bandwidth to b_{KDE} times the reference-score NRD0 bandwidth, where $b_{\mathrm{KDE}} \in \{0.75, 1, 1.5, 2, 3\}$. All other conditions remain fixed, with a target level of $\alpha = 0.15$. Results are reported in Table 7.

Across the evaluated range, most configurations maintain low empirical FDR, indicating that risk control is reasonably stable under bandwidth changes. Increasing the bandwidth reduces power in some configurations, suggesting that stronger smoothing can weaken the distinction between low-risk samples and the rest of the mixture. Both combined scores maintain high power across bandwidths, while the default setting already balances contamination control and clean-sample retention in most configurations.

Table 7: Effect of the shared bandwidth on selection at $\alpha = 0.15$. For each score, each cell reports the maximum empirical FDR and the minimum–maximum Power range across the nine model–dataset configurations. The default bandwidth multiplier is $b_{\mathrm{KDE}} = 1.5$. The maximum empirical FDR and the endpoints of the Power range may come from different configurations.

Score	$b_{\mathrm{KDE}} = 0.75$	$b_{\mathrm{KDE}} = 1$	$\boldsymbol{b_{\mathrm{KDE}} = 1.5}$	$b_{\mathrm{KDE}} = 2$	$b_{\mathrm{KDE}} = 3$
PPL	0.10; 0.64–1.00	0.09; 0.64–1.00	0.11; 0.60–1.00	0.12; 0.55–1.00	0.12; 0.44–1.00
Lowercase	0.16; 0.83–0.97	0.16; 0.84–0.98	0.16; 0.85–0.99	0.16; 0.83–0.99	0.17; 0.78–0.99
Zlib	0.11; 0.90–0.99	0.11; 0.90–0.99	0.12; 0.90–1.00	0.12; 0.89–1.00	0.13; 0.84–1.00
Grad	0.15; 0.63–0.95	0.15; 0.63–0.97	0.13; 0.63–0.98	0.11; 0.60–0.98	0.07; 0.51–0.97
Min-K 20%	0.14; 0.82–1.00	0.15; 0.83–1.00	0.15; 0.82–1.00	0.14; 0.81–1.00	0.14; 0.74–1.00
Min-K 30%	0.14; 0.87–1.00	0.15; 0.88–1.00	0.16; 0.88–1.00	0.16; 0.86–1.00	0.14; 0.81–1.00
PAC	0.14; 0.73–1.00	0.15; 0.74–1.00	0.15; 0.74–1.00	0.15; 0.72–1.00	0.14; 0.65–1.00
MEAN_Z	0.14; 0.91–1.00	0.15; 0.92–1.00	0.15; 0.92–1.00	0.15; 0.92–1.00	0.15; 0.89–1.00
First_PCA	0.15; 0.91–1.00	0.15; 0.92–1.00	0.15; 0.92–1.00	0.15; 0.92–1.00	0.14; 0.89–1.00

E.2 Distributional Separation and Risk Estimation

We use Gaussian simulations to examine distributional separation and the effects of proportion and density estimation on selection. This simulation examines performance under distributional overlap; Appendix E.3 examines the theoretical expectations under the stated conditions. Clean and contaminated scores follow $\mathcal{N}(0, 1)$ and $\mathcal{N}(-d, 1)$, respectively, with $d \in \{0.5, 1, 2, 3\}$ and reference sizes $m \in \{100, 500, 1000\}$. The candidate set contains 1,000 samples, with 500 clean and 500 contaminated samples, and each condition is repeated 200 times. Alongside CleanRef-LFDR, we substitute the true proportion, true densities, or true local risks to examine the contribution of each estimation step.

Distributional separation affects the trade-off between risk and power. With greater separation between the clean and contaminated distributions, CleanRef-LFDR achieves lower empirical FDR and substantially higher power. At $d = 3$, the method meets the risk target at every reference size, and power approaches complete recovery as the reference size increases. These results show that sufficient distributional information can support effective risk estimation, allowing the method to retain more clean samples while controlling residual contamination. Diagnostics under greater distributional overlap further illustrate the effect of proportion estimation on selection.

Proportion estimation affects risk assessment. When the distributions overlap substantially, substituting the true proportion markedly reduces empirical FDR, whereas replacing only the densities does not reduce empirical FDR. This suggests that the risk discrepancy in these settings is primarily associated with proportion estimation. At the same time, even with true local risks, few clean samples can be retained when separation is weak. Accurate risk estimation and informative score distributions therefore jointly determine clean-sample recovery under risk control.

E.3 Simulation under the Theoretical Conditions

To examine finite-sample performance under the theoretical conditions, we construct clean and contaminated densities as

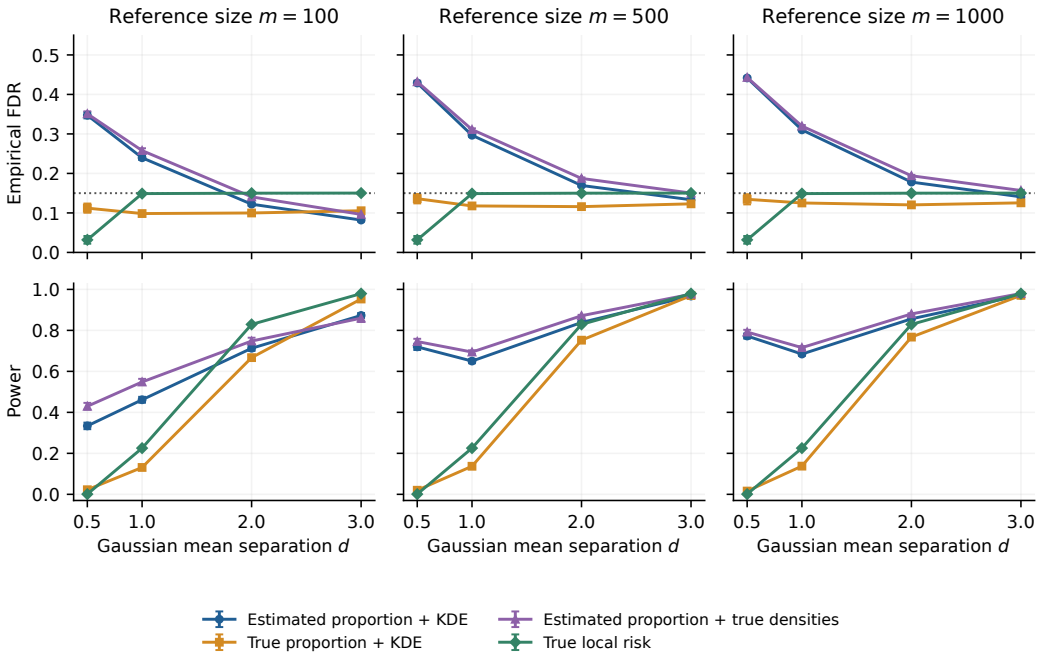

Figure 13: Effects of distributional separation and risk estimation. Columns correspond to reference sizes, and the two rows show empirical FDR and Power. Each point is the mean over 200 trials, with one Monte Carlo standard error. The dotted line denotes the target level $\alpha = 0.15$. The true-proportion, true-density, and true-local-risk variants examine the corresponding estimation steps.

Table 8: Complete Gaussian simulation results. Each cell reports empirical FDR / Power over 200 trials. All variants use the same generated samples.

		Selection methods		Paired diagnostics		
d	m	PU-BH	CleanRef-LFDR	True π	True densities	True risk
0.5	100	0.477 / 0.343	0.347 / 0.334	0.112 / 0.022	0.351 / 0.430	0.032 / 0.001
0.5	500	0.483 / 0.619	0.429 / 0.720	0.136 / 0.020	0.432 / 0.746	0.032 / 0.001
0.5	1000	0.484 / 0.575	0.442 / 0.772	0.134 / 0.016	0.443 / 0.792	0.032 / 0.001
1	100	0.416 / 0.371	0.240 / 0.461	0.098 / 0.131	0.258 / 0.548	0.149 / 0.225
1	500	0.447 / 0.746	0.297 / 0.650	0.118 / 0.137	0.311 / 0.694	0.149 / 0.225
1	1000	0.438 / 0.618	0.310 / 0.684	0.125 / 0.138	0.320 / 0.716	0.149 / 0.225
2	100	0.253 / 0.440	0.122 / 0.713	0.100 / 0.667	0.141 / 0.748	0.150 / 0.829
2	500	0.324 / 0.865	0.170 / 0.839	0.116 / 0.752	0.187 / 0.871	0.150 / 0.829
2	1000	0.349 / 0.890	0.178 / 0.856	0.120 / 0.767	0.194 / 0.880	0.150 / 0.829
3	100	0.116 / 0.495	0.082 / 0.873	0.106 / 0.953	0.096 / 0.860	0.150 / 0.979
3	500	0.173 / 0.905	0.133 / 0.967	0.123 / 0.970	0.150 / 0.977	0.150 / 0.979
3	1000	0.213 / 0.967	0.141 / 0.972	0.126 / 0.971	0.157 / 0.980	0.150 / 0.979

$$f_0(w) = b(w), \qquad f_1(w) = b(w+s), \qquad b(w) = 140w^3(1-w)^3\mathbf{1}\{0 < w < 1\},$$

where $s \in \{0.4, 0.8\}$ and the clean proportion is $\pi = 0.5$. These densities are smooth and compactly supported, with an interval containing only the clean component, satisfying the stated identifiability and quantitative separation conditions. The reference and mixture samples are generated independently, with sample sizes growing jointly as $m = n = N \in \{250, 500, 1000, 2000\}$. The shared bandwidth is $h_N = 0.25N^{-1/5}$, and each condition is repeated 200 times.

Figure 14 shows that, as the reference and mixture sample sizes increase together, the power of CleanRef-LFDR improves and approaches that of selection using true local risks, while empirical FDR remains near or below the target. With greater separation, even a small sample size supports high power, and additional samples bring power closer to one. These results illustrate how adequate reference information and distributional separation help the method balance contamination-risk control with clean-sample recovery.

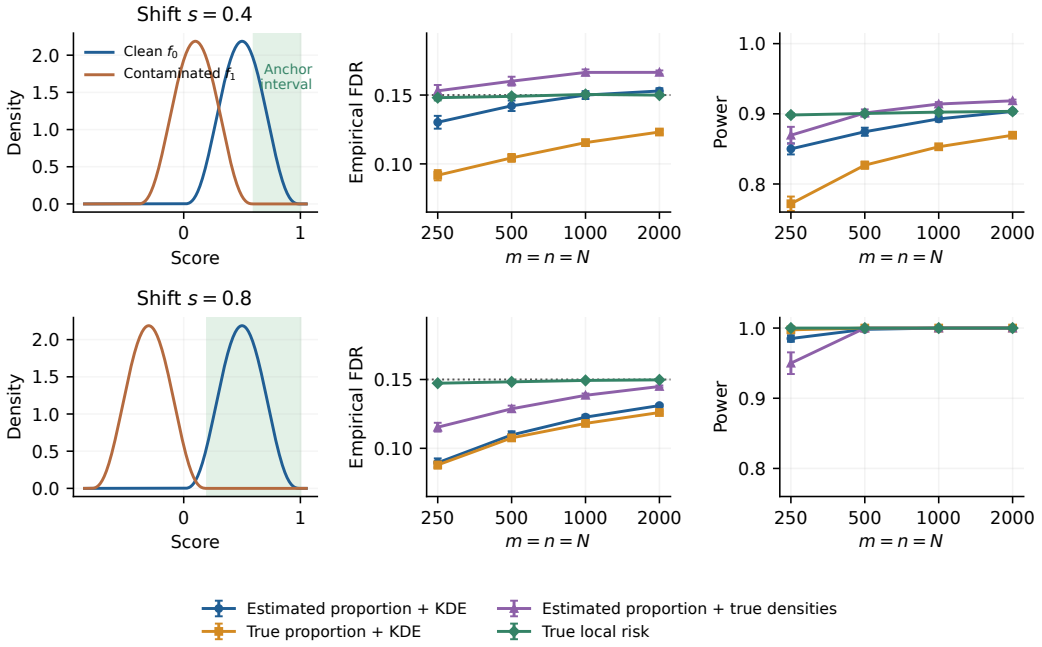

Figure 14: Finite-sample performance under the theoretical conditions. Rows correspond to $s = 0.4$ and $s = 0.8$, and columns show the component densities, empirical FDR, and Power. Each point is the mean over 200 independent trials, with one Monte Carlo standard error. The dotted line denotes the target level $\alpha = 0.15$.

F Application Details

F.1 Experimental Setup

Following the application study of FTD (Zhang et al., 2026), we examine how evaluation-set selection affects model rankings on SimpleQA (Wei et al., 2024) and ARC-C (Clark et al., 2018). We introduce contamination by fine-tuning Mistral-7B-Instruct-v0.3 on designated questions. The candidate evaluation set contains equal numbers of clean and contaminated questions, with partitions reported in Table 9. Each method then selects questions from the candidate set, and the models are re-evaluated on the selected questions.

Contamination fine-tuning uses LoRA (Hu et al., 2022) rank $r_{\mathrm{LoRA}} = 16$, scaling parameter $\alpha_{\mathrm{LoRA}} = 32$, and dropout 0.10. Training runs for 10 epochs with a batch size of 8, a learning rate of 5×10^{-5}, and a cosine learning-rate schedule.

The eight ranked models are Mistral-7B-Instruct-v0.3, Meta-Llama-3-70B-Instruct, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Llama-3.2-3B-Instruct, GPT-3.5-Turbo-0125, GPT-3.5-Turbo-1106, GPT-3.5-Turbo-Instruct, and GPT-4.1-Nano-2025-04-14. After contamination injection, only the model state of Mistral is replaced; the other models remain fixed.

Methods may retain different question subsets. We use the pre-injection ranking on the unfiltered candidate evaluation set as a common reference for ranking consistency after selection. Model answers are graded uniformly by GPT-4o-mini-2024-07-18. Within each leaderboard, all models are evaluated on the same questions and ranked by accuracy, with average ranks assigned to ties. We report Spearman's rank correlation coefficient (SRCC) (Sedgwick, 2014), together with FDP and Power on the fixed split, at a target level of $\alpha = 0.15$. An SRCC closer to one indicates greater agreement between the ranking after selection and the pre-injection ranking.

Table 9: Data partitions and question counts for the application study. The candidate evaluation sets contain equal numbers of clean and contaminated questions on both benchmarks.

Benchmark	Clean continuation	$\lvert\mathcal{D}_{\mathrm{cal}}\rvert$	Clean candidates	Contaminated candidates	$\lvert\mathcal{D}_{\mathrm{test}}\rvert$
SimpleQA	432	649	1082	1082	2164
ARC-C	117	176	293	293	586

F.2 Complete Results and Model Rankings

Table 10 summarizes selection and ranking results for all nine scoring configurations. Tables 12–17 provide detailed model rankings for First_PCA, Min-K 30%, and Min-K 20%. The main observations are as follows.

Selection improves ranking consistency. After selection with CleanRef-LFDR, rankings on both benchmarks are closer to the pre-injection rankings. On SimpleQA, most scoring configurations recover the complete pre-injection ordering; on ARC-C, every scoring configuration improves ranking consistency. Min-K 30%, shown in the main text, keeps FDP below the target and matches the best baseline SRCC on both benchmarks. These results show that the method effectively mitigates the impact of contamination on model rankings.

Higher power preserves more usable evaluation questions. On both benchmarks, CleanRef-LFDR retains more clean questions than both baselines for every scoring configuration. On ARC-C, all scoring configurations keep FDP below the target, and most achieve near-perfect power. These results show that selection based on local contamination risk can limit residual contamination while preserving sufficient clean evaluation data to support reliable model comparisons.

To examine the effect of changing the question subset, we also compare pre- and post-injection model rankings on the same retained questions. With Min-K 30%, CleanRef-LFDR yields identical orderings on SimpleQA and high ranking consistency on ARC-C. This comparison complements the results referenced to the unfiltered candidate evaluation set.

These comparisons address questions distinct from the SRCC reported against the common reference in the main text and are presented as supplementary results.

Model rankings after selection using the specified scoring configuration. Ranks are determined by accuracy. SRCC measures the agreement between each ranking column and the pre-injection ranking.

Table 10: Complete application results for nine scoring configurations at $\alpha = 0.15$, reporting FDP, Power, and SRCC. Each method evaluates models on its retained questions. SRCC is measured against the pre-injection ranking on the unfiltered candidate evaluation set.

	Raw			fsd			CleanRef-LFDR		
Score	FDP	Power	SRCC	FDP	Power	SRCC	FDP	Power	SRCC
SimpleQA									
PPL	0.284	0.762	0.976	0.329	0.730	0.976	0.158	0.800	1.000
Lowercase	0.253	0.802	0.976	0.216	0.787	0.976	0.185	0.955	1.000
Zlib	0.306	0.768	0.976	0.206	0.832	0.970	0.152	0.930	1.000
Grad	0.166	0.789	1.000	0.207	0.773	0.976	0.095	0.929	1.000
Min-K 20%	0.323	0.726	0.976	0.218	0.751	0.976	0.095	0.910	0.976
Min-K 30%	0.286	0.765	0.952	0.188	0.824	1.000	0.142	0.945	1.000
PAC	0.337	0.743	0.976	0.286	0.760	0.976	0.120	0.839	0.976
MEAN_Z	0.244	0.763	0.976	0.149	0.831	1.000	0.155	0.965	1.000
First_PCA	0.244	0.780	0.976	0.164	0.789	1.000	0.152	0.961	1.000
ARC-C									
PPL	0.137	0.840	0.976	0.105	0.843	0.952	0.149	0.976	0.952
Lowercase	0.180	0.826	0.976	0.080	0.829	0.952	0.080	0.976	0.952
Zlib	0.155	0.819	0.952	0.039	0.833	0.952	0.119	0.990	0.952
Grad	0.077	0.778	0.952	0.059	0.877	0.929	0.004	0.884	0.952
Min-K 20%	0.094	0.860	0.952	0.044	0.809	0.952	0.101	1.000	0.952
Min-K 30%	0.098	0.877	0.952	0.038	0.874	0.946	0.107	1.000	0.952
PAC	0.116	0.860	0.994	0.078	0.843	0.952	0.068	0.990	0.952
MEAN_Z	0.066	0.819	0.952	0.033	0.799	0.952	0.039	1.000	0.952
First_PCA	0.079	0.833	0.976	0.052	0.805	0.952	0.087	1.000	0.952

Table 11: Ranking comparisons on the same question subset. Results use Min-K 30%, CleanRef-LFDR, and a target level of $\alpha = 0.15$. Both columns report SRCC.

Benchmark	Pre-injection ranking on retained questions vs pre-injection ranking on the full candidate set	Post-injection vs pre-injection ranking on the same retained questions
SimpleQA	1.000	1.000
ARC-C	0.976	0.976

Table 12: SimpleQA rankings with First_PCA at $\alpha = 0.15$.

Pre-injection	Unfiltered	Raw	fsd	CleanRef-LFDR
1 Llama-3-70B	1.5 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B
2 GPT-3.5-1106	**1.5 Mistral-7B**	2 GPT-3.5-1106	2 GPT-3.5-1106	2 GPT-3.5-1106
3 GPT-3.5-0125	3 GPT-3.5-1106	3 GPT-3.5-0125	3 GPT-3.5-0125	3 GPT-3.5-0125
4 GPT-3.5-Inst.	4 GPT-3.5-0125	4 GPT-3.5-Inst.	4 GPT-3.5-Inst.	4 GPT-3.5-Inst.
5 GPT-4.1-Nano	5 GPT-3.5-Inst.	**5 Mistral-7B**	5 GPT-4.1-Nano	5 GPT-4.1-Nano
6 Mistral-7B	6 GPT-4.1-Nano	6 GPT-4.1-Nano	**6 Mistral-7B**	**6 Mistral-7B**
7 Llama-3.1-8B	7 Llama-3.1-8B	7 Llama-3.1-8B	7 Llama-3.1-8B	7 Llama-3.1-8B
8 Llama-3.2-3B	8 Llama-3.2-3B	8 Llama-3.2-3B	8 Llama-3.2-3B	8 Llama-3.2-3B
SRCC 1.000	SRCC 0.707	SRCC 0.976	SRCC 1.000	SRCC 1.000
FDP –	FDP –	FDP 0.244	FDP 0.164	FDP 0.152

Table 13: SimpleQA rankings with Min-K 30% at $\alpha = 0.15$.

Pre-injection	Unfiltered	Raw	fsd	CleanRef-LFDR
1 Llama-3-70B	1.5 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B
2 GPT-3.5-1106	**1.5 Mistral-7B**	2 GPT-3.5-0125	2 GPT-3.5-1106	2 GPT-3.5-1106
3 GPT-3.5-0125	3 GPT-3.5-1106	3 GPT-3.5-1106	3 GPT-3.5-0125	3 GPT-3.5-0125
4 GPT-3.5-Inst.	4 GPT-3.5-0125	4 GPT-3.5-Inst.	4 GPT-3.5-Inst.	4 GPT-3.5-Inst.
5 GPT-4.1-Nano	5 GPT-3.5-Inst.	**5 Mistral-7B**	5 GPT-4.1-Nano	5 GPT-4.1-Nano
6 Mistral-7B	6 GPT-4.1-Nano	6 GPT-4.1-Nano	**6 Mistral-7B**	**6 Mistral-7B**
7 Llama-3.1-8B	7 Llama-3.1-8B	7 Llama-3.1-8B	7 Llama-3.1-8B	7 Llama-3.1-8B
8 Llama-3.2-3B	8 Llama-3.2-3B	8 Llama-3.2-3B	8 Llama-3.2-3B	8 Llama-3.2-3B
SRCC 1.000	SRCC 0.707	SRCC 0.952	SRCC 1.000	SRCC 1.000
FDP –	FDP –	FDP 0.286	FDP 0.188	FDP 0.142

Table 14: SimpleQA rankings with Min-K 20% at $\alpha = 0.15$.

Pre-injection	Unfiltered	Raw	fsd	CleanRef-LFDR
1 Llama-3-70B	1.5 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B
2 GPT-3.5-1106	**1.5 Mistral-7B**	2 GPT-3.5-1106	2 GPT-3.5-1106	2 GPT-3.5-1106
3 GPT-3.5-0125	3 GPT-3.5-1106	3 GPT-3.5-0125	3 GPT-3.5-0125	3 GPT-3.5-0125
4 GPT-3.5-Inst.	4 GPT-3.5-0125	4 GPT-3.5-Inst.	4 GPT-3.5-Inst.	4 GPT-3.5-Inst.
5 GPT-4.1-Nano	5 GPT-3.5-Inst.	**5 Mistral-7B**	**5 Mistral-7B**	5 GPT-4.1-Nano
6 Mistral-7B	6 GPT-4.1-Nano	6 GPT-4.1-Nano	6 GPT-4.1-Nano	6 Llama-3.1-8B
7 Llama-3.1-8B	7 Llama-3.1-8B	7 Llama-3.1-8B	7 Llama-3.1-8B	**7 Mistral-7B**
8 Llama-3.2-3B	8 Llama-3.2-3B	8 Llama-3.2-3B	8 Llama-3.2-3B	8 Llama-3.2-3B
SRCC 1.000	SRCC 0.707	SRCC 0.976	SRCC 0.976	SRCC 0.976
FDP –	FDP –	FDP 0.323	FDP 0.218	FDP 0.095

Table 15: ARC-C rankings with First_PCA at $\alpha = 0.15$.

Pre-injection	Unfiltered	Raw	fsd	CleanRef-LFDR
1 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B
2 GPT-4.1-Nano	2 GPT-4.1-Nano	2 GPT-4.1-Nano	2 GPT-4.1-Nano	2 GPT-4.1-Nano
3 GPT-3.5-0125	3 GPT-3.5-0125	3 GPT-3.5-0125	3 GPT-3.5-0125	3 GPT-3.5-0125
4 GPT-3.5-1106	4 GPT-3.5-1106	4 GPT-3.5-1106	4 GPT-3.5-1106	4 GPT-3.5-1106
5 GPT-3.5-Inst.	**5 Mistral-7B**	5 GPT-3.5-Inst.	5 Llama-3.1-8B	5 Llama-3.1-8B
6 Llama-3.1-8B	6 GPT-3.5-Inst.	6 Llama-3.1-8B	6 GPT-3.5-Inst.	6 GPT-3.5-Inst.
7 Mistral-7B	7 Llama-3.1-8B	7 Llama-3.2-3B	7 Llama-3.2-3B	7 Llama-3.2-3B
8 Llama-3.2-3B	8 Llama-3.2-3B	**8 Mistral-7B**	**8 Mistral-7B**	**8 Mistral-7B**
SRCC 1.000	SRCC 0.929	SRCC 0.976	SRCC 0.952	SRCC 0.952
FDP –	FDP –	FDP 0.079	FDP 0.052	FDP 0.087

Table 16: ARC-C rankings with Min-K 30% at $\alpha = 0.15$.

Pre-injection	Unfiltered	Raw	fsd	CleanRef-LFDR
1 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B	1.5 GPT-4.1-Nano	1 Llama-3-70B
2 GPT-4.1-Nano	2 GPT-4.1-Nano	2 GPT-4.1-Nano	1.5 Llama-3-70B	2 GPT-4.1-Nano
3 GPT-3.5-0125	3 GPT-3.5-0125	3 GPT-3.5-0125	3 GPT-3.5-0125	3 GPT-3.5-0125
4 GPT-3.5-1106	4 GPT-3.5-1106	4 GPT-3.5-1106	4 GPT-3.5-1106	4 GPT-3.5-1106
5 GPT-3.5-Inst.	**5 Mistral-7B**	5 Llama-3.1-8B	5 Llama-3.1-8B	5 Llama-3.1-8B
6 Llama-3.1-8B	6 GPT-3.5-Inst.	6 GPT-3.5-Inst.	6 GPT-3.5-Inst.	6 GPT-3.5-Inst.
7 Mistral-7B	7 Llama-3.1-8B	7 Llama-3.2-3B	7 Llama-3.2-3B	7 Llama-3.2-3B
8 Llama-3.2-3B	8 Llama-3.2-3B	**8 Mistral-7B**	**8 Mistral-7B**	**8 Mistral-7B**
SRCC 1.000	SRCC 0.929	SRCC 0.952	SRCC 0.946	SRCC 0.952
FDP –	FDP –	FDP 0.098	FDP 0.038	FDP 0.107

Table 17: ARC-C rankings with Min-K 20% at $\alpha = 0.15$.

Pre-injection	Unfiltered	Raw	fsd	CleanRef-LFDR
1 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B	1 Llama-3-70B
2 GPT-4.1-Nano	2 GPT-4.1-Nano	2 GPT-4.1-Nano	2 GPT-4.1-Nano	2 GPT-4.1-Nano
3 GPT-3.5-0125	3 GPT-3.5-0125	3 GPT-3.5-0125	3 GPT-3.5-0125	3 GPT-3.5-0125
4 GPT-3.5-1106	4 GPT-3.5-1106	4 GPT-3.5-1106	4 GPT-3.5-1106	4 GPT-3.5-1106
5 GPT-3.5-Inst.	**5 Mistral-7B**	5 Llama-3.1-8B	5 Llama-3.1-8B	5 Llama-3.1-8B
6 Llama-3.1-8B	6 GPT-3.5-Inst.	6 GPT-3.5-Inst.	6 GPT-3.5-Inst.	6 GPT-3.5-Inst.
7 Mistral-7B	7 Llama-3.1-8B	7 Llama-3.2-3B	7 Llama-3.2-3B	7 Llama-3.2-3B
8 Llama-3.2-3B	8 Llama-3.2-3B	**8 Mistral-7B**	**8 Mistral-7B**	**8 Mistral-7B**
SRCC 1.000	SRCC 0.929	SRCC 0.952	SRCC 0.952	SRCC 0.952
FDP –	FDP –	FDP 0.094	FDP 0.044	FDP 0.101