Controllable Contamination Detection for Reliable LLM Evaluation with Statistical Guarantees

Zheng Zhang[1], Qi Liu[1]*, Siyuan Liang[2], Ning Li[1], Zirui Hu[2], Weibo Gao[1], Rui Li[1], Zhenya Huang[1], Leszek Rutkowski[3], Baosheng Yu[2], Dacheng Tao[2]

[1]State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China

[2]Nanyang Technological University [3]The AGH University of Krakow

zhangzheng@mail.ustc.edu.cn, qiliuql@ustc.edu.cn, siyuan.liang@ntu.edu.sg, ningli03@mail.ustc.edu.cn

zirui.hu@ntu.edu.sg, {weibogao, ruili2000}@mail.ustc.edu.cn, huangzhy@ustc.edu.cn

rutkowski@agh.edu.pl, {baosheng.yu, dacheng.tao}@ntu.edu.sg

Abstract

Large language models (LLMs) have achieved remarkable performance across diverse tasks, largely driven by large-scale pretraining. However, this data abundance introduces test data contamination, where benchmark datasets overlap with pretraining corpora, undermining the reliability of model evaluation by confounding memorization with genuine generalization. To mitigate this issue, existing training data detectors attempt to identify clean (unseen) samples from contaminated test sets, but often suffer from residual contamination due to the black-box nature of LLMs. As a result, contaminated data may be mistakenly retained, leading to unreliable evaluation. To address this challenge, we propose **FTD** (**F**DR-controlled **T**raining **D**ata detection), a principled framework that detects and filters contaminated evaluation data while providing a statistical guarantee: the proportion of contaminated samples mistakenly retained as clean, the false discovery rate (FDR), is provably controlled below a user-specified threshold. FTD combines multiple complementary detectors via an adaptive weighting strategy, and we theoretically show it achieves high statistical power under valid FDR control. Extensive experiments on real-world benchmarks demonstrate that FTD significantly reduces residual contamination compared to existing methods while preserving evaluation consistency.

*Corresponding author.

1 Introduction

Large language models (LLMs) have achieved remarkable performance across many domains (Zhao et al., 2023; Gao et al., 2025; Zhan et al., 2025; Zhang et al., 2025b; Mao et al., 2024; Ye et al., 2026), largely enabled by training on massive-scale datasets (Cheng et al., 2025; Chang et al., 2024). However, the uncontrolled expansion of such data has also led to test data contamination, where evaluation benchmarks are inadvertently included in the training corpus. This contamination blurs the distinction between genuine generalization and memorization, thereby undermining the reliability of LLM evaluation (Zhang et al., 2024c; Lv et al., 2024; Yao et al., 2024; Sun et al., 2025), which is critical in high-stakes domains such as education (Wang et al., 2024a; Gao et al., 2023).

To mitigate this issue, numerous training data detection methods have been proposed (Shi et al., 2024; Zhang et al., 2024b,a; Golchin and Surdeanu, 2023; Jacovi et al., 2023). These approaches typically cast the problem as a binary classification task, aiming to separate contaminated samples (seen during training) from clean, unseen data within a potentially contaminated test set. The identified clean data can then be used for more reliable evaluation of LLMs.

However, for most deployed LLMs, training data, model internals, and gradients are inaccessible, leaving detectors to rely solely on observable signals such as output probabilities or per-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30122–30143

plexity. As a result, even state-of-the-art detection methods cannot perfectly distinguish contaminated data from clean samples. Consequently, a non-negligible fraction of contaminated data may be misclassified as clean (i.e., false positives), leading to residual contamination in the selected evaluation set and undermining evaluation reliability (as shown in Section 5.4). If the proportion of such false positives can be controlled below a user-specified threshold, residual contamination can be effectively mitigated. From a statistical perspective, this problem naturally lends itself to false discovery rate (FDR) control. Classical procedures such as the Benjamini–Hochberg (BH) method (Benjamini and Hochberg, 1995) provide rigorous guarantees and have been extensively studied (Benjamini and Yekutieli, 2001). Nevertheless, existing FDR-controlled approaches often overlook statistical power, i.e., retaining as many truly clean samples as possible. Insufficient power may result in overly conservative filtering, yielding too few clean samples and compromising the representativeness of subsequent evaluation. This motivates our primary objective: can we achieve the FDR control while maximizing statistical power for training data detection?

To settle this problem, we propose FTD, a training data detection framework for reliable LLM evaluation that jointly targets FDR control and power maximization. FTD introduces a principled fusion strategy that integrates multiple complementary detectors, including PPL (Li, 2023), Min-k (Shi et al., 2024), and Min-k++ (Zhang et al., 2024b), within the classical BH framework. Our approach proceeds in three steps. First, for each candidate sample, we extract test statistics from individual detectors and compute the corresponding p-values. Second, since different detectors may perform differently across datasets, we propose an adaptive weighting strategy based on each detector's detection performance, which learns the relative contribution of each detector. Using these learned weights, we then combine multiple detectors via a weighted combination. Finally, the BH procedure is applied to the combined p-values to achieve valid FDR control.

From a theoretical perspective, to account for data heterogeneity, we employ an adaptive weighting scheme rather than fixed weights. Since these weights are estimated from the same samples used to compute p-values, they are stochastically dependent on the underlying test statistics, which renders existing theoretical guarantees for fixed-weight combinations inapplicable (Bates et al., 2023; Wu et al., 2023; Long et al., 2023). To overcome this difficulty, we establish the convergence of the data-driven weights and extend the analysis to the resulting converged regime. Our theoretical results show that FTD achieves asymptotic optimality, simultaneously ensuring valid FDR control and high statistical power.

Finally, we evaluate FTD on multiple real-world datasets, including WikiMIA (Shi et al., 2024), arXivTection (Duarte et al., 2024), BBC RealTime (Li et al., 2024), and MIMIR (Duan et al., 2024). In our experiments, existing methods frequently fail to control the FDR below the 0.15 threshold (e.g., Zlib reaches an FDR as high as 0.292 and Lowercase up to 0.216), indicating that a substantial fraction of contaminated samples can still pass through. In contrast, FTD consistently reduces the FDR to below 0.101 across all settings, demonstrating its strong ability to control false positives. In addition, we assess FTD in real-world benchmark evaluations by constructing contaminated settings, where LLMs are fine-tuned on four classical evaluation benchmarks. The results show that FTD consistently identifies clean data under controlled FDR, effectively mitigating the impact of contamination while preserving the reliable LLM evaluation.

- To the best of our knowledge, we are the first to propose controlling the FDR while maximizing power specifically for clean-data selection, a setting tailored to reliable LLM evaluation. This effectively mitigates the impact of contaminated data and enhances the reliability of LLM evaluation.
- We introduce FTD, a training data detection method that offers theoretical guarantees for FDR control and power maximization.
- We validate the effectiveness of our approach on real-world datasets and demonstrate its capability to support reliable model evaluation.

2 Related Works

Data Contamination. Trustworthy machine learning has attracted increasing attention in recent years (Zhang et al., 2023, 2025c; Liu et al., 2025; Liang et al., 2025; Ying et al., 2026; Ren et al., 2025; Liang et al., 2026; Hu et al., 2026), with data contamination emerging as a

critical challenge. Data contamination has been extensively studied in the literature (Mann et al., 2020; Magar and Schwartz, 2022; Deng et al., 2024; Golchin and Surdeanu, 2024a; Xu and Yan, 2025; Zhang et al., 2025a; Bordt et al., 2025; Kocyigit et al., 2025), where training data may inadvertently include evaluation benchmark data, leading to unreliable evaluation results. As such, assessing the potential leakage of benchmark data into pretraining corpora is essential for trustworthy model evaluation (Dong et al., 2024; Dekoninck et al., 2024; Zhao et al., 2025a,b; Oren et al., 2023; Golchin and Surdeanu, 2024b). In fact, the problem of data contamination can, to some extent, be viewed as a specific instance of Membership Inference Attacks (MIA) (Duan et al., 2024), which aim to determine whether a given data point was part of a models training set. Motivated by this connection, several recent works have approached training data detection in LLMs from the perspective of MIA. For example, (Shi et al., 2024) hypothesize that clean data is more likely to contain outlier tokens that result in significantly higher loss values. Based on this observation, they propose the Min-K% method, which identifies contaminated data by analyzing the top-k token log probabilities.

FDR Control. Our training data detection problem can be naturally formulated as a multiple testing problem, where the goal is to retain a subset of samples that are declared clean for downstream evaluation. In this setting, a key reliability criterion is the false discovery rate (FDR), namely, the expected proportion of contaminated samples among those retained as clean. Controlling FDR is a classical objective in statistics and has been widely studied as a principled way to balance reliable discovery with error control (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001), with important applications in areas such as genomics (Storey and Tibshirani, 2003) and healthcare (Genovese et al., 2002). More recently, multiple testing has also received growing attention in the machine learning literature, where it has been used as a principled tool for reliable outlier detection and sample selection (Bates et al., 2023; Wang et al., 2024b; Wu et al., 2024).

For training data detection, however, FDR control alone is not sufficient. One must also maintain adequate statistical power to recover as many truly clean samples as possible, since residual contamination in the retained evaluation set can substantially distort the reliability of LLM assessment. This makes it essential to jointly study error control and power, especially when combining multiple detectors through data-driven weights.

Several prior works use related statistical tools but pursue different goals. Oren et al. (2023) use permutation tests to establish the existence of test-set contamination, but their objective is contamination detection at the dataset level rather than selecting a clean subset for evaluation. Dekoninck et al. (2024) identify contamination through anomalous performance patterns, but do not provide FDR guarantees for the subset retained as clean. The closest work is Hu et al. (2025), which also casts training data detection as a multiple testing problem with FDR control. However, their goal is to identify contaminated samples for security and attribution, whereas our goal is to retain clean samples for reliable evaluation. As a result, the null and alternative hypotheses are reversed: in their setting, discoveries correspond to contaminated samples, while in ours, discoveries correspond to clean samples.

In contrast, we propose FTD to support rigorous and reliable LLM evaluation by simultaneously controlling FDR and improving statistical power on the retained clean set. Moreover, our method incorporates adaptive multi-detector fusion with data-driven weights, and we establish theoretical guarantees that explicitly account for the endogenous dependence between the learned weights and the resulting p-values.

3 Problem Definitions

Suppose we are given a dataset consisting of n samples, denoted by $Z_1, Z_2, \ldots, Z_n$, with index set $\mathcal{D}_{\text{total}} = [n] = \{1, 2, \ldots, n\}$. Each sample is either contaminated or clean. Specifically, a sample is called **contaminated** if it has been seen during the training process of the LLM; otherwise, it is called **clean**. Accordingly, we define two disjoint index sets:

$$\mathcal{D}_{\text{con}} \subseteq \mathcal{D}_{\text{total}}, \qquad \mathcal{D}_{\text{clean}} = \mathcal{D}_{\text{total}} \setminus \mathcal{D}_{\text{con}},$$

where $\mathcal{D}_{\text{con}}$ and $\mathcal{D}_{\text{clean}}$ denote the contaminated and clean samples, respectively.

Our goal is to select a subset of samples that can be declared clean and then used for reliable LLM evaluation. We denote this selected subset by $\hat{\mathcal{S}} \subseteq \mathcal{D}_{\text{total}}$. The quality of $\hat{\mathcal{S}}$ is assessed by two

criteria: (i) its reliability, measured by the proportion of contaminated samples mistakenly retained in $\hat{\mathcal{S}}$, and (ii) its efficiency, measured by how many truly clean samples are successfully retained.

To construct $\hat{\mathcal{S}}$, each sample Z_i is assigned a detection score S_i, where lower scores generally indicate a higher likelihood of contamination. A thresholding rule can then be used to determine whether a sample is declared contaminated or clean:

$$h(Z_i) = \begin{cases} \text{contaminated data}, & \text{if } S_i < t, \\ \text{clean data}, & \text{if } S_i \geq t, \end{cases} \quad (1)$$

where t is a threshold determined from the score distribution.

Many existing detectors produce such scores, for example, Min-K%, which uses the average probability of the $k\%$ outlier tokens with the lowest predicted probabilities. However, because LLMs are largely black-box systems—with inaccessible pretraining corpora and limited model transparency—training data detection remains challenging. As a result, even strong detection methods may still retain a non-negligible number of contaminated samples, which undermines the reliability of downstream evaluation.

To address this issue, we adopt a multiple testing perspective. For each sample $i \in \mathcal{D}_{\text{total}}$, we consider the hypothesis test

$$\begin{cases} \mathbb{H}_0 : \text{Sample } i \text{ is contaminated}, \\ \mathbb{H}_1 : \text{Sample } i \text{ is clean}. \end{cases} \quad (2)$$

A rejection of $\mathbb{H}_0$ means that sample i is selected into $\hat{\mathcal{S}}$ and declared clean. Under this formulation, our objective is to control the false discovery rate of $\hat{\mathcal{S}}$ below a user-specified level α, while maximizing power. These two quantities are defined as

$$\begin{aligned} \text{FDR} &:= \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{D}_{\text{con}}|}{|\hat{\mathcal{S}}| \vee 1}\right], \\ \text{Power} &:= \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{D}_{\text{clean}}|}{|\mathcal{D}_{\text{clean}}| \vee 1}\right]. \end{aligned} \quad (3)$$

That is, FDR quantifies the residual contamination among the samples we retain as clean, while power measures the fraction of truly clean samples successfully preserved for evaluation.

4 Methods

In this section, we introduce FTD, a method for detecting non-training data with rigorous FDR control and power maximization, enabling reliable evaluation of LLMs. Specifically, we first formalize the training data detection task as a multiple hypothesis testing problem and review the classical BH (Benjamini and Hochberg, 1995) procedure for FDR control. Building upon BH, we then propose a Cauchy fusion that integrates existing training data detection methods to maximize statistical power. Finally, we provide theoretical guarantees for the FTD.

4.1 The BH Procedure

To control the FDR, we formulate the training data detection task as a multiple hypothesis testing problem. Specifically, for each test sample Z_i, we test the null hypothesis,

$H_0^{(i)}$: "Sample Z_i is from the contaminated data".

By computing a p-value p_i for each sample, we obtain a collection of n hypotheses $\{H_0^{(i)}\}_{i=1}^n$, one for each test sample. This formulation enables the application of classical multiple testing procedures to determine which hypotheses to reject.

To obtain the p-value, we first compute a detection score S_i for each sample Z_i using a given detection method. Let $\mathcal{S}_{\text{ref}}$ denote the contaminated data detection score distribution. Although the true distribution of contaminated data is not directly accessible, it can be approximated by collecting samples from commonly used pretraining sources, such as Wikipedia, that are from the same domain and were published prior to the models release date. For example, WIKIMIA (Shi et al., 2024) considers Wikipedia articles created before 2017 as contaminated data, because many pretrained models, including LLaMA and GPT-NeoX, were released after 2017 and incorporate Wikipedia dumps into their pretraining corpora. The p-value for sample Z_i is then defined as:

$$p_i = \mathbb{P}_{S \sim \mathcal{S}_{\text{ref}}}(S \leq S_i). \quad (4)$$

The BH procedure is a classical method in multiple hypothesis testing, which aims to control the error rate by identifying a data-dependent threshold such that FDR is bounded by a target level $\alpha \in (0, 1)$. Specifically, given a set of p-values $\{p_i\}_{i=1}^n$, we first sort them in ascending order: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$, and then find the largest index:

$$m = \max\left\{j \in [n] : p_{(j)} \leq \frac{j\alpha}{n}\right\}, \quad (5)$$

The resulting BH threshold is then given by $t_{\text{BH}} = \frac{m\alpha}{n}$. Finally, all p-values smaller than t_{BH}

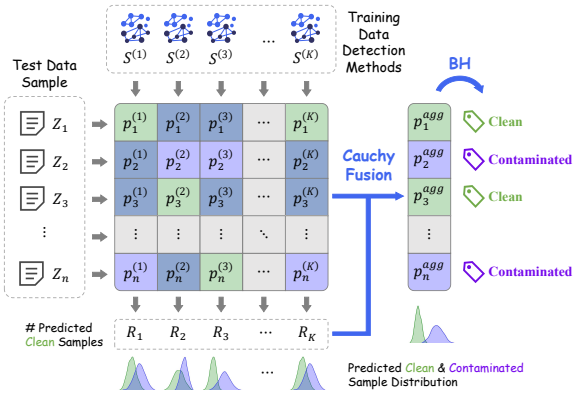

Figure 1: The framework of FTD.

are considered statistically significant, leading to the rejection of the null hypothesis H_0, and the corresponding samples are identified as clean data. The final $\widehat{\mathcal{S}}$ can be represent as:

$$\widehat{\mathcal{S}} = \{Z_j \in \mathcal{D}_{\text{total}} : p_j < t_{\text{BH}}\}. \tag{6}$$

4.2 FTD

While the BH procedure effectively controls the FDR, it overlooks the goal of maximizing statistical power. This can result in only a small number of clean samples being detected, thereby compromising the comprehensiveness of model evaluation. Recent advances in training data detection have introduced various scoring methods (e.g., Min-K (Shi et al., 2024), Min-K++ (Zhang et al., 2024b), etc.), each capturing different aspects of the data. These methods provide complementary information, which motivates the idea of combining multiple statistics to construct a more powerful p-value. The framework is shown in Figure 1. Specifically, FTD proceeds in three stages: **Stage 1 – Adaptive Weight Learning**: for each detector $S^{(k)}$, a weight w_k is learned from its BH rejection count, reflecting its relative detection performance; **Stage 2 – Weighted Cauchy P-value Aggregation**: the per-detector p-values $p_i^{(k)}$ are combined into a single aggregated p-value p_i^{agg} via a weighted Cauchy transform; **Stage 3 – FDR Control via BH**: BH is applied to $\{p_i^{\text{agg}}\}$ at level α to produce the predicted-clean set $\hat{\mathcal{S}}$.

Stage 1: Adaptive Weight Learning. Suppose we have a set of training data detection methods, each associated with a score function $S^{(1)}, \ldots, S^{(K)}$. The corresponding p-values based on these score functions are denoted as $\{p_j^{(1)}\}_{j\in[n]}, \ldots, \{p_j^{(K)}\}_{j\in[n]}$, respectively. Since the performance of each detection method may vary, we aim to better integrate the information from different models by assigning them appropriate weights. Intuitively, the more non-training data points a method successfully identifies (i.e., rejects the null hypothesis H_0), the more it contributes to the overall power (Zhang et al., 2022b). Therefore, we compute the weight of each method $S^{(k)}$ based on the number of rejections it makes under a controlled FDR:

$$\begin{aligned} R_k &= \left|\left\{Z_i \in \mathcal{D}_{\text{total}} : p_i^{(k)} \le t_{\text{BH}}^{(k)}\right\}\right|, \\ w_k &= \frac{R_k}{\sum_{j=1}^{K} R_j}. \end{aligned} \tag{7}$$

Stage 2: Weighted Cauchy P-value Aggregation. Here, $t_{\text{BH}}^{(k)}$ is the rejection threshold determined by the BH procedure over $\{p_j^{(k)}\}_{j\in[n]}$. Inspired by (Liu and Xie, 2020), we map the individual p-values to the Cauchy space and compute a weighted sum:

$$T_i = \sum_{k=1}^{K} w_k \cdot \tan\left[\left(0.5 - p_i^{(k)}\right)\pi\right]. \tag{8}$$

The reason why we choose the Cauchy distribution is that it is a heavy-tailed distribution. This implies that when an individual p-value is very small (i.e., highly significant), its corresponding Cauchy-transformed value becomes extremely large (approaching infinity). Such behavior allows the combined statistic T_i to be dominated by the most significant evidence among the individual tests, thereby increasing the sensitivity to strong signals while maintaining robustness under H_0. The combined test statistic T_i approximately follows a standard Cauchy distribution. To map it back to the p-value space, we apply the inverse Cauchy transformation, yielding the aggregated p-value:

$$p_i^{\text{agg}} = \frac{1}{2} - \frac{1}{\pi}\arctan(T_i). \tag{9}$$

Stage 3: FDR Control via BH. The aggregated p-value p_i^{agg} is then used in the BH procedure to control the overall FDR and maximize power. The final threshold t is defined as:

$$t_{final} = \frac{\alpha}{n}\max\left\{j \in [n] : p_{(j)}^{\text{agg}} \le \frac{j\alpha}{n}\right\}, \tag{10}$$

where $p_{(j)}^{\text{agg}}$ denotes the jth smallest value among the set $\{p_i^{\text{agg}}\}_{i\in[n]}$. The predicted-clean set is then:

$$\widehat{\mathcal{S}} = \{Z_j \in \mathcal{D}_{\text{total}} : p_j < t_{final}\}. \tag{11}$$

4.3 Theoretical Guarantee

In this section, we theoretically prove that FTD can effectively control the FDR while maximizing statistical power. We begin by presenting two key lemmas, which form the theoretical foundation for establishing the FDR control guarantee of the FTD procedure. Specifically, we first characterize the statistical properties of the aggregated p-values used in FTD. Based on these properties, we then demonstrate the effectiveness of FTD in controlling the FDR.

Lemma 1. *(Bates et al., 2023) Under the null hypothesis H_0, suppose the score $S_i^{(k)}$ is drawn from the same distribution as the reference data, i.e., $S_i^{(k)} \sim \mathcal{S}_{ref}^{(k)}$, and the p-value is defined as $p_i^{(k)} = \mathbb{P}_{S \sim \mathcal{S}_{ref}^{(k)}}(S \leq S_i^{(k)})$. Then, under H_0, the p-value $p_i^{(k)}$ follows a uniform distribution: $p_i^{(k)} \sim \mathrm{Uniform}[0,1]$.*

Lemma 1 has been proved by (Bates et al., 2023). Building on Lemma 1, consider the p-values $p_i^{(1)}, \ldots, p_i^{(K)}$ computed from different training data detectors. Each of these p-values satisfies,

$$p_i^{(k)} \sim \mathrm{Uniform}[0,1] \quad \text{under } H_0.$$

Under fixed (deterministic) weights, prior work establishes that the Cauchy-aggregated p-values $\{p_i^{\mathrm{agg}}\}_{i\in[n]}$ are approximately uniformly distributed under H_0 (Liu and Xie, 2020), a result that has been extended to various dependence structures (Wu et al., 2023; Long et al., 2023). However, in our method the weights are data-driven (computed from BH rejection counts), so they are stochastically dependent on the same p-values used in the aggregation. This dependence means the fixed-weight validity results cannot be directly applied. To address this, we proceed in two steps: we first show in Lemma 2 that the data-driven weights converge almost surely to deterministic constants as $n \to \infty$. By a continuous-mapping argument, this implies that the aggregated statistic converges to its fixed-weight counterpart, and Lemma 3 then establishes that the aggregated p-values are asymptotically uniform under H_0, enabling FDR control via BH (Theorem 1).

Lemma 2 (Convergence of the weights for FTD)**.** *Consider K detection methods, each with score function $S^{(1)}, \ldots, S^{(K)}$. For each method k, let p-values on the total dataset $\mathcal{D}_{\mathrm{total}}$ be $\{p_i^{(k)}\}_{i=1}^n$. Apply the BH procedure at level α separately to each method, producing the rejection threshold $t_{\mathrm{BH}}^{(k)}$ and the number of rejections $R_k = \left|\left\{Z_i \in \mathcal{D}_{\mathrm{total}} : p_i^{(k)} \leq t_{\mathrm{BH}}^{(k)}\right\}\right|$. Define the normalized data-driven weights $w_k = \frac{R_k}{\sum_{j=1}^K R_j}$. Then, there exist deterministic constants $(w_1^*, \ldots, w_K^*)$ with $\sum_{k=1}^K w_k^* = 1$ such that $w_k \xrightarrow{\text{a.s.}} w_k^*$ for each $k = 1, \ldots, K$.*

Lemma 3 (Uniformity of aggregated p-values in FTD)**.** *Let $p_i^{(1)}, \ldots, p_i^{(K)}$ be the p-values corresponding to a given sample i, each satisfying $p_i^{(k)} \sim \mathrm{Uniform}[0,1]$ under the null hypothesis H_0. We have that the aggregated p-value p_i^{agg} is uniformly distributed on $[0,1]$ under H_0.*

Remark 1. *A detailed theoretical analysis supporting Lemmas 2 and 3 is provided in Appendix A . In addition, we empirically evaluate the distribution of aggregated p-values p_i^{agg} under the null hypothesis on real datasets. The results show that the aggregated p-values are uniformly distributed on $[0,1]$, which provides empirical evidence supporting the validity of Lemma 3.*

Based on the result of Lemma 3, we have the following theorems:

Theorem 1 (FDR Control of FTD)**.** *Suppose we are given a set of training data detection methods, each associated with a score function $S^{(1)}, \ldots, S^{(K)}$. Let the corresponding p-values be defined as $\{p_j^{(1)}\}_{j\in[n]}, \ldots, \{p_j^{(K)}\}_{j\in[n]}$, where each p-value is computed as $p_i^{(k)} = \mathbb{P}_{S \sim \mathcal{S}_{ref}^{(k)}}(S \leq S_i^{(k)})$. Then, the FTD procedure controls the FDR at level α, i.e.,*

$$FDR := \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{D}_{con}|}{|\hat{\mathcal{S}}| \vee 1}\right] \leq \alpha. \tag{12}$$

Theorem 2 (Asymptotic Power Consistency of FTD)**.** *Assume that for each $k \in [K]$ and $Z_j \in \mathcal{D}_{\mathrm{clean}}$, $\Pr(p_j^{(k)} \leq c) \geq 1 - \delta$ where $\delta = o(\sqrt{\log n / n})$ and c is a constant, and the density function of p_j^{agg} for $Z_j \in \mathcal{D}_{\mathrm{clean}}$ has an upper bound $C_f > 0$, we have,*

$$Power := \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{D}_{clean}|}{|\mathcal{D}_{clean}| \vee 1}\right] \geq 1 - C\sqrt{\log n / n}. \tag{13}$$

The assumption in Theorem 2 is standard for proving power consistency (Genovese et al., 2002;

Method	WikiMIA									arXivTection									BBC Real Time								
	Pythia-2.8B			OPT-6.7B			LLaMA-13B			Pythia-2.8B			OPT-6.7B			LLaMA-13B			Pythia-2.8B			OPT-6.7B			LLaMA-13B		
	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑
PPL	0.141	0.980	0.913	0.168	0.821	0.835	0.145	1.000	0.919	0.149	0.909	0.873	0.156	0.839	0.839	0.138	0.960	0.901	0.156	**0.802**	0.825	0.192	0.604	0.727	0.137	0.931	0.890
Lowercase	0.154	0.926	0.884	0.208	0.671	0.758	0.149	0.999	0.916	0.216	0.615	0.718	0.212	0.588	0.710	0.154	0.832	0.837	0.173	0.694	0.772	0.190	**0.617**	**0.733**	0.133	0.928	0.892
Zlib	0.178	0.768	0.809	0.244	0.545	0.698	0.150	0.959	0.899	0.200	0.580	0.712	0.243	0.449	0.646	0.160	0.819	0.829	0.207	0.550	0.700	0.292	0.356	0.600	0.151	0.832	0.840
Grad	0.174	0.798	0.823	0.190	0.717	0.784	0.154	0.929	0.885	0.150	0.798	0.825	0.167	0.732	0.789	0.139	0.913	0.881	0.188	0.645	0.745	0.205	0.586	0.714	0.151	0.853	0.849
Min-K%	0.150	0.982	0.908	0.167	0.895	0.864	0.151	1.000	0.915	0.135	**0.962**	0.904	0.136	**0.898**	0.877	0.141	0.995	0.915	0.187	0.656	0.750	0.221	0.513	0.680	0.138	0.910	0.881
Min-K%++	0.145	0.997	0.918	0.153	**0.974**	0.903	0.150	1.000	0.915	0.148	0.906	0.872	0.154	0.812	0.829	0.137	0.995	0.917	0.167	0.744	0.795	0.202	0.595	0.719	0.133	0.935	0.894
FTD	**0.101**	**0.999**	**0.946**	**0.100**	0.964	**0.931**	**0.099**	**1.000**	**0.947**	**0.090**	0.942	**0.923**	**0.083**	0.849	**0.884**	**0.092**	**0.997**	**0.947**	**0.095**	0.771	**0.843**	**0.098**	0.506	0.722	**0.094**	**0.948**	**0.924**
(std)	±0.018	±0.001	±0.011	±0.017	±0.012	±0.006	±0.012	±0.000	±0.007	±0.014	±0.010	±0.007	±0.021	±0.030	±0.006	±0.020	±0.002	±0.012	±0.013	±0.017	±0.003	±0.006	±0.037	±0.016	±0.005	±0.002	±0.003

Table 1: The detection performance on three datasets. FDR (lower is better), Power (higher is better), ACC (higher is better). Bold indicates the best result per model. For FTD, we report the mean and standard deviation across ten independent runs.

Weinstein et al., 2023). Specifically, we show that the power of FTD is lower bounded by $1 - C\sqrt{\log n/n}$, where $C > 0$ is a constant depending on the testing level, the null proportion, and certain distributional characteristics. Consequently, as $n \to \infty$, the power converges to 1. Through Theorems 1 and 2, we demonstrate that FTD achieves asymptotic optimality in controlling the FDR while maintaining high statistical power. Details will be presented in Appendix B.

5 Experiments

In this section, we present the experimental results to demonstrate the effectiveness of our model. The code is publicly available at `https://github.com/liano3/FTD`.

5.1 Setup

Baselines We select several baselines from prior work on training data detection, including PPL (Li, 2023), Lowercase (Carlini et al., 2021), Zlib (Carlini et al., 2021), Grad (Hu et al., 2025), Min-K% (Shi et al., 2024), and Min-K%++ (Zhang et al., 2024b). In addition, to assess the effectiveness of our adaptive weighting strategy, we construct three FTD variants: BH-Average, BH-Random, and BH-Max. Details are in Appendix C.1.

Models and Datasets We adopt three LLMs to evaluate our FTD: Pythia-2.8B (Biderman et al., 2023), OPT-6.7B (Zhang et al., 2022a) and LLaMA-13B (Touvron et al., 2023). As for the evaluation datasets, we employ four benchmark datasets for evaluations, including WikiMIA (Shi et al., 2024), ArXivTection (Duarte et al., 2024), BBC Real Time (Li et al., 2024), MIMIR (Duan et al., 2024). Details are in Appendix C.2.

Evaluation metrics In this paper, we focus on FDR and Power (Eq. (3)), which play a crucial role in training data detection for LLM evaluation. A lower FDR and a higher Power indicate better detection performance. Additionally, we report the Accuracy (ACC), where a higher ACC reflects more precise detection results.

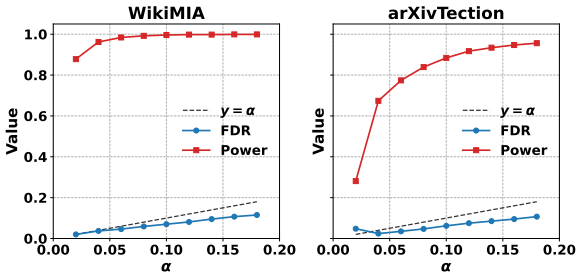

Figure 2: The impact of α on Pythia-2.8B.

Implementation details The detailed implementation can be found in in Appendix C.3.

5.2 Main Result

In this section, we conduct a comprehensive comparison of various detection methods across three datasets and multiple language models. Table 1 summarizes the detection performance of all evaluated methods across three datasets and under multiple LLM backbones. Results for MIMIR are provided in the Appendix C.4. The key findings are as follows: ❶ *FTD achieves the lowest FDR across all datasets and model settings, demonstrating its strong ability to suppress false positives.* Compared to the strongest baseline, FTD achieves a relative FDR reduction of up to 30–39%, indicating a substantial improvement in precision when identifying clean data. ❷ *FTD maintains competitive or superior statistical power in most settings.* It achieves near-perfect power on two datasets and performs robustly on the third, showing that it can effectively recall clean samples while maintaining low FDR. ❸*FTD consistently achieves the highest overall accuracy across different datasets and models.* This reflects its comprehensive effectiveness in both minimizing false detections and correctly identifying clean data.

5.3 The Analysis of FTD

Effect of Cauchy Fusion. We adopt the weighted Cauchy fusion strategy defined in Eq. (8)

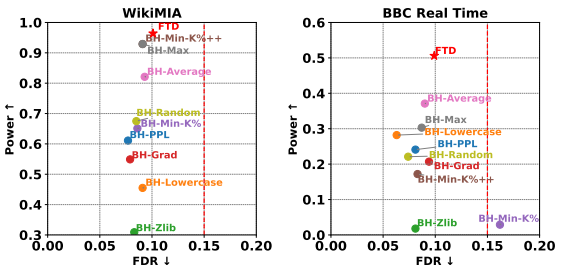

Figure 3: Ablation study results on Pythia-2.8B.

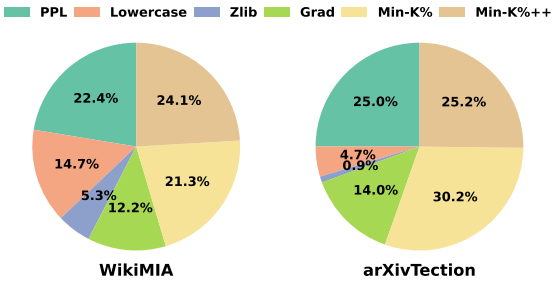

Figure 4: Learned weights for each method in the FTD on Pythia-2.8B.

and Eq. (9) and evaluate its effectiveness from two aspects. First, to assess the benefit of fusion, we remove the fusion module and directly apply the BH procedure (Section 4.1) to p-values from individual detectors, yielding baselines denoted as BH-XX. Second, to examine the impact of the weighting scheme, we consider three variants: BH-Average, BH-Random, and BH-Max. Figure 3 reports the FDR and Power across datasets. We observe that ❶*all BH-based methods control the FDR near the 0.15 threshold in the majority of settings*, demonstrating the robustness of BH under both individual and fused p-values. Notably, certain curves (e.g., BH-Min-K%) occasionally exceed 0.15 in individual runs, which is also reasonable: the FDR guarantee is an *expectation-type* bound on the false discovery proportion (FDP) averaged over randomness, rather than a per-run guarantee; the realized FDP in a single experiment can therefore transiently exceed the nominal level due to sampling variability; and ❷*FTD achieves the highest Power while maintaining valid FDR control*, outperforming BH-Average, BH-Random, and BH-Max, which highlights the advantage of the learned weighting scheme.

Effect of α. To investigate the impact of the FDR control threshold α on model performance, we vary α from 0.02 to 0.18 and compute the corresponding FDR and Power for each method. Each subplot in Figure 2 presents results on a different dataset, and each curve corresponds to a specific metric. To clearly visualize the FDR constraint, we also include a dashed line representing the target threshold ($y = \alpha$). A model is considered to successfully control the FDR if its corresponding FDR curve remains below this line. As α increases, the observed FDR (blue curve) consistently remains below the target threshold (dashed line $y = \alpha$) across all datasets, demonstrating that FTD successfully controls the FDR within the specified bounds. This indicates that the FTD is statistically valid and reliable across a wide range of α values.

Method	Simple QA				GPQA			
	LLaMA-3.1-8B		Mistral-7B		LLaMA-3.1-8B		Mistral-7B	
	FDR↓	SRCC↑	FDR↓	SRCC↑	FDR↓	SRCC↑	FDR↓	SRCC↑
Contaminated	-	0.657	-	0.500	-	0.143	-	0.886
PPL	0.134	0.886	0.223	0.810	0.056	0.771	0.148	0.829
Lowercase	0.135	0.829	0.248	0.833	0.045	0.829	0.113	0.943
Zlib	0.125	0.829	0.234	0.833	0.086	0.600	0.227	0.714
Grad	0.116	**1.000**	0.108	0.929	0.056	0.600	0.179	0.886
Min-K%	0.139	0.943	0.199	0.810	0.054	0.657	0.142	0.886
Min-K%++	0.118	**1.000**	0.172	0.857	0.031	0.829	0.105	0.943
FTD	**0.096**	**1.000**	**0.025**	**1.000**	**0.025**	**0.943**	**0.057**	**1.000**

Table 2: Impact of different detection methods on LLM Evaluation for SimpleQA and GPQA. SRCC measures whether a detector can effectively remove contaminated data and thereby preserve ranking consistency in model evaluation; values closer to 1 indicate higher consistency. The **Contaminated** row reports the benchmark results of the contaminated models without applying any training data detection methods.

Anaysis of learning weight w_i. To further examine the effectiveness of our weighted fusion strategy, we visualize the learned weights assigned to each method across different datasets, as shown in Figure 4. The pie charts correspond to the datasets WikiMIA, arXivTection. Each segment represents the weight of a base method in the final fusion. We observe that methods with stronger performance (as reported in the table 1) tend to receive larger weights. For instance, Min-K%++ and Min-K% consistently receive higher weights across all datasets, while Grad and Zlib are assigned smaller weights. This alignment between performance and weighting indicates that our fusion mechanism is capable of distinguishing the power of each method, and integrating them in a power-aware manner.

5.4 The Application of FTD

In this section, we explore the practical application of FTD to assess its ability to mitigate the impact

of contamination on LLM evaluation. Since existing data contamination benchmarks do not support LLM evaluation, following (Hu et al., 2025), we construct a synthetic contaminated setting. Specifically, we fine-tune an LLM on a subset of the evaluation data to intentionally introduce contamination. Training data detection methods are then applied to identify clean samples and conduct evaluation. Because different detection methods rely on different scoring signals (e.g., likelihood, compression ratio, or token-level statistics) and thus flag different samples as contaminated, they retain different clean benchmark subsets. As a result, directly comparing absolute performance scores across methods would be unfair. We therefore focus on **ranking consistency**, arguing that a reliable detection method should preserve the relative ranking of models observed on clean benchmarks.

More concretely, we consider four widely used LLM evaluation benchmarks, SimpleQA, GPQA, TruthfulQA, and ARC-C, with detailed descriptions provided in Appendix D.1. We evaluate eight models on these benchmarks, and the resulting rankings under clean conditions are treated as the ground truth. To simulate contamination, we randomly select 50% of the benchmark data to fine-tune two models (LLaMA-3.1-8B and Mistral-7B), thereby introducing artificial contamination; the detailed fine-tuning setup is provided in Appendix D.2. We then apply training data detection methods to filter contaminated samples and re-evaluate all models. The closer the resulting rankings are to the clean-condition ground truth, the more effective the detection method. For quantitative evaluation, we adopt Spearmans Rank Correlation Coefficient (SRCC) (Sedgwick, 2014) as the consistency metric, where values closer to 1 indicate better preservation of reliable evaluation. Results on SimpleQA and GPQA are shown in Table 2, and additional results on TruthfulQA and ARC-C are deferred to Appendix D.3.

From the table, we draw the following observations. ❶ *Data contamination severely compromises fair model evaluation*. After contamination is introduced, the rankings of affected models deviate substantially from the clean-condition ground truth, leading to misleading conclusions about their true capabilities. ❷ *Existing training data detection methods are helpful but insufficient*. While they partially mitigate the impact of contamination, their performance remains unsatisfactory in terms of both FDR and SRCC. ❸ *FTD effectively mitigates data contamination in model evaluation*. FTD consistently achieves the lowest FDR among all methods, and its SRCC reaches 1 across all benchmarks, highlighting that strict FDR control is essential for minimizing contamination effects and ensuring reliable LLM evaluation.

We further evaluate the effectiveness of FTD under different contamination levels. Detailed results are provided in Appendix D.4, which further demonstrate that FTD yields reliable LLM evaluation across varying contamination ratios.

6 Conclusion

In this work, we studied test data contamination in LLM evaluation and proposed FTD, a principled framework that achieves strict FDR control while preserving clean evaluation data. By combining multiple complementary detectors via an adaptive fusion strategy with a statistical testing procedure, FTD provides both theoretical guarantees and strong empirical performance. Experiments on real-world datasets show that FTD consistently outperforms state-of-the-art methods, underscoring the value of statistically grounded approaches for reliable LLM evaluation.

7 Limitations

In this section, we discuss the limitations of FTD. While FTD demonstrates strong performance in controlling the FDR and maximizing statistical power, certain limitations remain. Specifically, to balance FDR control with high statistical power, we introduce a fusion strategy that integrates multiple detection techniques into the classical BH procedure. However, the effectiveness of this framework depends heavily on the quality of the underlying detection methods. If the base detectors (e.g., Min-K%) fail to capture specific contamination patterns, the aggregated detection signal may remain weak, potentially resulting in undetected contaminated samples. In future work, we aim to develop more powerful detection methods to further mitigate the impact of test data contamination.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (No.62525606, No.62477044), the Key Technologies R & D Program of Anhui Province (No. 202423k09020039), and the Young Elite Scientists Sponsorship Program by CAST (No. 2024QNRC001).

References

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

Yoav Benjamini and Yosef Hochberg. 1995. 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.

Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle OBrien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sebastian Bordt, Suraj Srinivas, Valentyn Boreiko, and Ulrike von Luxburg. 2025. How much can we forget about data contamination? In *Forty-second International Conference on Machine Learning*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

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

Yuxing Cheng, Yi Chang, and Yuan Wu. 2025. A survey on data contamination for large language models. *arXiv preprint arXiv:2502.14425*.

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

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

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. 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)*, pages 8698–8711.

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

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? In *First Conference on Language Modeling*.

André V Duarte, Xuandong Zhao, Arlindo L Oliveira, and Lei Li. 2024. De-cop: Detecting copyrighted content in language models training data. *arXiv preprint arXiv:2402.09910*.

William Feller. 1991. *An introduction to probability theory and its applications, Volume 2*, volume 2. John Wiley & Sons.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. 2025. Agent4edu: Generating learner response data by generative agents for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 22, pages 23923–23932.

Weibo Gao, Hao Wang, Qi Liu, Fei Wang, Xin Lin, Linan Yue, Zheng Zhang, Rui Lv, and Shijin Wang. 2023. Leveraging transferable knowledge concept graph embedding for cold-start cognitive diagnosis. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 983–992.

Christopher R Genovese, Nicole A Lazar, and Thomas Nichols. 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878.

Shahriar Golchin and Mihai Surdeanu. 2023. Data contamination quiz: A tool to detect and estimate contamination in large language models. *arXiv preprint arXiv:2311.06233*.

Shahriar Golchin and Mihai Surdeanu. 2024a. Time travel in llms: Tracing data contamination in large language models. In *ICLR*.

Shahriar Golchin and Mihai Surdeanu. 2024b. Time travel in llms: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*.

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

Zirui Hu, Zheng Zhang, Yingjie Wang, Leszek Rutkowski, and Dacheng Tao. 2026. Cofact: Conformal factuality guarantees for language models under covariate shift. In *The Fourteenth International Conference on Learning Representations*.

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

Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 2025. Overestimation in llm evaluation: A controlled large-scale study on data contaminations impact on machine translation. In *Forty-second International Conference on Machine Learning*.

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

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. 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, pages 18600–18607.

Siyuan Liang, Tianmeng Fang, Zhe Liu, Aishan Liu, Yan Xiao, Jinyuan He, Ee-Chien Chang, and Xiaochun Cao. 2025. Safemobile: Chain-level jailbreak detection and automated evaluation for multimodal mobile agents. *arXiv preprint arXiv:2507.00841*.

Siyuan Liang, Jiajun Gong, Tianmeng Fang, Aishan Liu, Tao Wang, Xiaochun Cao, Dacheng Tao, and Chang Ee-Chien. 2026. Trapflow: Controllable website fingerprinting defense via dynamic backdoor learning. *IEEE Transactions on Information Forensics and Security*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Yaowu Liu and Jun Xie. 2020. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 504–514.

Mingya Long, Zhengbang Li, Wei Zhang, and Qizhai Li. 2023. The cauchy combination test under arbitrary dependence structures. *The American Statistician*, 77(2):134–142.

Wenxi Lv, Qinliang Su, Hai Wan, Hongteng Xu, and Wenchao Xu. 2024. Contamination-resilient anomaly detection via adversarial learning on partially-observed normal and anomalous data. In *Forty-first International Conference on Machine Learning*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*.

Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, and 1 others. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1:3.

Qingyang Mao, Qi Liu, Zhi Li, Mingyue Cheng, Zheng Zhang, and Rui Li. 2024. Potable: Towards systematic thinking via stage-oriented plan-then-execute reasoning on tables. *arXiv preprint arXiv:2412.04272*.

Yonatan Oren, Nicole Meister, Niladri S Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2023. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *Preprint*, arXiv:2311.12022.

Zhiyao Ren, Siyuan Liang, Aishan Liu, and Dacheng Tao. 2025. Iclshield: Exploring and mitigating in-context learning backdoor attacks. *arXiv preprint arXiv:2507.01321*.

Philip Sedgwick. 2014. Spearmans rank correlation coefficient. *Bmj*, 349.

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

John D Storey, Jonathan E Taylor, and David Siegmund. 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal*

of the Royal Statistical Society Series B: Statistical Methodology, 66(1):187–205.

John D Storey and Robert Tibshirani. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

Yifan Sun, Han Wang, Dongbai Li, Gang Wang, and Huan Zhang. 2025. The emperor's new clothes in benchmarking? a rigorous examination of mitigation strategies for llm benchmark data contamination. In *Forty-second International Conference on Machine Learning*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Fei Wang, Weibo Gao, Qi Liu, Jiatong Li, Guanhao Zhao, Zheng Zhang, Zhenya Huang, Mengxiao Zhu, Shijin Wang, Wei Tong, and 1 others. 2024a. A survey of models for cognitive diagnosis: New developments and future directions. *arXiv preprint arXiv:2407.05458*.

Xiaoning Wang, Yuyang Huo, Liuhua Peng, and Changliang Zou. 2024b. Conformalized multiple testing after data-dependent selection. *Advances in Neural Information Processing Systems*, 37:58574–58609.

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

Asaf Weinstein, Weijie J Su, Małgorzata Bogdan, Rina Foygel Barber, and Emmanuel J Candes. 2023. A power analysis for model-x knockoffs with ℓ_p-regularized statistics. *The Annals of Statistics*, 51(3):1005–1029.

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

Xiaoyang Wu, Yuyang Huo, and Changliang Zou. 2023. Multi-split conformal prediction via cauchy aggregation. *Stat*, 12(1):e522.

Cheng Xu and Nan Yan. 2025. Triplefact: Defending data contamination in the evaluation of llm-driven fake news detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8808–8823.

Feng Yao, Yufan Zhuang, Zihao Sun, Sunan Xu, Animesh Kumar, and Jingbo Shang. 2024. Data contamination can cross language barriers. *arXiv preprint arXiv:2406.13236*.

Shengyu Ye, Qi Liu, Hao Jiang, Zheng Zhang, Heng Yu, and Zhenya Huang. 2026. Themis: Automated constraint-aware test synthesis framework for code reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34432–34440.

Zonghao Ying, Aishan Liu, Siyuan Liang, Lei Huang, Jinyang Guo, Wenbo Zhou, Xianglong Liu, and Dacheng Tao. 2026. Safebench: A safety evaluation framework for multimodal large language models. *International Journal of Computer Vision*, 134(1):18.

Yi Zhan, Qi Liu, Weibo Gao, Zheng Zhang, Tianfu Wang, Shuanghong Shen, Junyu Lu, and Zhenya Huang. 2025. Coderagent: Simulating student behavior for personalized programming learning with large language models. *arXiv preprint arXiv:2505.20642*.

Hengxiang Zhang, Songxin Zhang, Bingyi Jing, and Hongxin Wei. 2024a. Fine-tuning can help detect pretraining data from large language models. *arXiv preprint arXiv:2410.10880*.

Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. 2025a. Position: Membership inference attacks cannot prove that a model was trained on your data. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 333–345. IEEE.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2024b. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*.

Kun Zhang, Jingyu Li, Zhe Li, and S Kevin Zhou. 2025b. Dh-set: Improving vision-language alignment with diverse and hybrid set-embeddings learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24993–25003.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022a. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten Rijke, Yixing Fan, and Xueqi Cheng. 2024c. Pretraining data detection for large language models: A divergence-based calibration method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5263–5274.

Yifan Zhang, Haiyan Jiang, Haojie Ren, Changliang Zou, and Dejing Dou. 2022b. Automs: automatic model selection for novelty detection with error rate control. *Advances in Neural Information Processing Systems*, 35:19917–19929.

Zheng Zhang, Ning Li, Qi Liu, Rui Li, Weibo Gao, Qingyang Mao, Zhenya Huang, Baosheng Yu, and

Dacheng Tao. 2025c. The other side of the coin: Exploring fairness in retrieval-augmented generation. *arXiv preprint arXiv:2504.12323*.

Zheng Zhang, Qi Liu, Hao Jiang, Fei Wang, Yan Zhuang, Le Wu, Weibo Gao, and Enhong Chen. 2023. Fairlisa: fair user modeling with limited sensitive attributes information. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 41432–41450.

Jingqian Zhao, Bingbing Wang, Geng Tu, Yice Zhang, Qianlong Wang, Bin Liang, Jing Li, and Ruifeng Xu. 2025a. Coreeval: Automatically building contamination-resilient datasets with real-world knowledge toward reliable llm evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22284–22306.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and 1 others. 2025b. Mmlu-cf: A contamination-free multi-task language understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13371–13391.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124.

A Appendix A

FTD is supported by strong statistical guarantees, enabling effective control of the FDR while maximizing statistical power. In this section, we present two key lemmas that form the theoretical foundation for establishing the FDR control guarantees of the FTD procedure. Specifically, we characterize the statistical properties of the aggregated p-values used in FTD from both empirical and theoretical perspectives.

Lemma 1. *(Bates et al., 2023) Under the null hypothesis H_0, suppose the score $S_i^{(k)}$ is drawn from the same distribution as the reference data, i.e., $S_i^{(k)} \sim \mathcal{S}_{ref}^{(k)}$, and the p-value is defined as $p_i^{(k)} = \mathbb{P}_{S\sim\mathcal{S}_{ref}^{(k)}}(S \leq S_i^{(k)})$. Then, under H_0, the p-value $p_i^{(k)}$ follows a uniform distribution: $p_i^{(k)} \sim \mathrm{Uniform}[0,1]$.*

This proof has already been established in (Bates et al., 2023); we refer the reader to that work for details. Then, we provide the detailed proof of Lemma 2. Before presenting the proof, we introduce the BH-limit lemma, which is a well-established result in classical multiple testing theory. It states that the ratio of rejections produced by the BH procedure to the total number of tests converges to a constant. This result is crucial for establishing the convergence of the data-driven weights.

Lemma (BH-limit (Storey et al., 2004))**.** *For each method $k \in \{1,\dots,K\}$, suppose the p-values $\{p_i^{(k)}\}_{i=1}^n$ are i.i.d. with a mixture distribution $F_k(t) = \Pr(p_i^{(k)} \leq t) = (1-\pi_k)t + \pi_k G_k(t), t \in [0,1]$, where $0 \leq \pi_k \leq 1$ and G_k is continuous. Let R_k denote the number of rejections from applying the BH procedure at level α to method k, and define the rejection proportion $r_{k,n} := \frac{R_k}{n}$. Then there exists a constant $r_k \in [0,1]$, such that, $r_{k,n} \xrightarrow{\text{a.s.}} r_k \quad (n \to \infty)$.*

Lemma 2 (Convergence of BH-based weights)**.** *Consider K detection methods, each with score function $S^{(1)},\dots,S^{(K)}$. For each method k, let p-values on the total dataset $\mathcal{D}_{\text{total}}$ be $\{p_i^{(k)}\}_{i=1}^n$. Apply the BH procedure at level α separately to each method, producing the rejection threshold $t_{\text{BH}}^{(k)}$ and the number of rejections $R_k = \left|\left\{Z_i \in \mathcal{D}_{\text{total}} : p_i^{(k)} \leq t_{\text{BH}}^{(k)}\right\}\right|$. Define the normalized data-driven weights $w_k = \frac{R_k}{\sum_{j=1}^K R_j}$. Then, there exist deterministic constants $(w_1^*,\dots,w_K^*)$ with $\sum_{k=1}^K w_k^* = 1$ such that $w_k \xrightarrow{\text{a.s.}} w_k^*$ for each $k = 1,\dots,K$.*

Proof. First, by Lemma BH-limit lemma, for each method k there exists a deterministic constant $r_k \in [0,1]$ such that the rejection proportion

$$r_{k,n} := \frac{R_k}{n} \xrightarrow{\text{a.s.}} r_k.$$

Since K is finite, the vector of rejection proportions $\mathbf{r}_n = (r_{1,n},\dots,r_{K,n})$ converges almost surely to $\mathbf{r} = (r_1,\dots,r_K)$.

Then, due to $\sum_{k=1}^K r_k > 0$, we define the limiting weights

$$w_k^* := \frac{r_k}{\sum_{j=1}^K r_j}, \quad k = 1,\dots,K.$$

The sum of the sample proportions $\sum_{k=1}^K r_{k,n}$ is eventually positive almost surely, so the weights w_k are well-defined for large n.

Finally, consider the continuous mapping

$$h : \{x \in \mathbb{R}^K : \sum_i x_i \neq 0\} \to \mathbb{R}^K,$$
$$h(x) = \Big(\frac{x_1}{\sum_i x_i}, \ldots, \frac{x_K}{\sum_i x_i}\Big).$$

which maps the rejection proportions to the weights. Applying the continuous mapping theorem to the almost-sure convergence of $\mathbf{r}_n$ gives

$$h(\mathbf{r}_n) = (w_1, \ldots, w_K) \xrightarrow{\text{a.s.}} h(\mathbf{r}) = (w_1^*, \ldots, w_K^*),$$

establishing the almost-sure convergence of the data-driven weights. Convergence in probability follows immediately. □

Base on Lemma 2, we can give the following lemma.

Lemma 3. *Let $p_i^{(1)}, \ldots, p_i^{(K)}$ be the p-values corresponding to a given sample i, each satisfying $p_i^{(k)} \sim \text{Uniform}[0,1]$ under the null hypothesis H_0. We have, the aggregated p-value p_i^{agg} is uniformly distributed on $[0,1]$ under H_0.W*

Proof. In this proof, we first consider the case where the weights are fixed, and then extend the result to the setting with data-driven weights. For simplicity, the fixed weights are denoted as w_k^* for each candidate $k = 1, \ldots, K$.

Let $U_k = p_i^{(k)} \sim \text{Uniform}[0,1]$. Define the transformed variable

$$X_k = \tan\left[\pi\left(0.5 - U_k\right)\right].$$

Now we prove that $X_k \sim \text{Cauchy}(0,1)$. To see this, recall that the cumulative distribution function (CDF) of the standard Cauchy distribution is

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x),$$

whose inverse is

$$F^{-1}(u) = \tan\left[\pi(u - 0.5)\right].$$

Therefore, if $U_k \sim \text{Uniform}[0,1]$, then,

$$X_k = \tan\left[\pi(0.5 - U_k)\right] = -\tan\left[\pi(U_k - 0.5)\right] \sim \text{Cauchy}(0,1).$$

since the Cauchy distribution is symmetric about zero.

Now, consider the weighted sum,

$$T^* = \sum_{k=1}^K w_k^* X_k.$$

Because the Cauchy distribution is stable under linear combinations (Feller, 1991), we have,

$$T^* \sim \text{Cauchy}(0,1).$$

Finally, we apply the inverse CDF of the standard Cauchy distribution to obtain the aggregated p-value,

$$p_i^{*,\text{agg}} = 1 - F(T_i^*) = \frac{1}{2} - \frac{1}{\pi}\arctan(T_i^*).$$

The distribution of $p_i^{*,\text{agg}}$ can be directly verified by Liu and Xie (2020) as a uniform distribution.

Next, we consider aggregated p-values based on data-driven weights. It suffices to prove the convergence between T and T^*. Denote

$$\Delta_k = w_k - w_k^*.$$

Then,

$$T - T^* = \sum_{k=1}^K (w_k - w_k^*)X_k = \sum_{k=1}^K \Delta_k X_k \leq \sum_{k=1}^K |\Delta_k|\,|X_k|.$$

Fix any k. Since X_k is standard Cauchy and $\Delta_k = o_p(1)$, note the linear scaling property of Cauchy distributions,

$$\Delta_k X_k \sim \text{Cauchy}(0, |\Delta_k|).$$

As $|\Delta_k| \to 0$ in probability, the scale parameter of $\Delta_k X_k$ converges to zero. By the definition of convergence in distribution, this implies

$$\Delta_k X_k \xrightarrow{p} 0,$$

i.e., each term is $o_p(1)$.

Since K is fixed, a finite sum of $o_p(1)$ terms is still $o_p(1)$,

$$\sum_{k=1}^K \Delta_k X_k = o_p(1).$$

Thus,

$$T - T^* = o_p(1).$$

Since T^* is a finite linear combination of independent standard Cauchy random variables, it is also Cauchy,

$$T^* = \sum_{k=1}^K w_k X_k \sim \text{Cauchy}\Big(0, \sum_{k=1}^K |w_k|\Big).$$

By Slutsky's theorem, adding a term that is $o_p(1)$ does not change the limiting distribution. Therefore,

$$T = T^* + o_p(1) \quad \Longrightarrow \quad T \xrightarrow{d} T^*.$$

This completes the proof. □

After establishing the theoretical foundation, we empirically examine the distribution of the aggregated p-values, p_i^{agg}. Specifically, we extract p_i^{agg} from two benchmark datasets, WikiMIA and arXivTection, and visualize their empirical distributions, as shown in Figure 5. The results indicate that the aggregated p-values approximately follow a uniform distribution, which provides empirical support for the validity of Lemma 3.

B Appendix B

In this section, we first theoretically demonstrate that FTD achieves asymptotic optimality in controlling the FDR while maintaining high statistical power, as established in Theorems 1 and 2.

Theorem 1 (FDR Control of FTD)**.** *Suppose we are given a set of training data detection methods, each associated with a score function $S^{(1)}, \ldots, S^{(K)}$. Let the corresponding p-values be defined as $\{p_j^{(1)}\}_{j\in[n]}, \ldots, \{p_j^{(K)}\}_{j\in[n]}$, where each p-value is computed as $p_i^{(k)} = \mathbb{P}_{S\sim\mathcal{S}_{ref}^{(k)}}(S \leq S_i^{(k)})$. Then, the FTD procedure controls the FDR at level α, i.e.,*

$$FDR := \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{D}_{con}|}{|\hat{\mathcal{S}}| \vee 1}\right] \leq \alpha. \tag{14}$$

Proof. From Lemma 3, the aggregated p-value p_i^{agg}, computed from $p_i^{(1)}, \ldots, p_i^{(K)}$, is approximately uniformly distributed on $[0, 1]$ under H_0. This satisfies the assumptions required by the BH procedure: namely, that p-values under H_0 are uniformly distributed.

According to the BH procedure (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001), if the p-values under the null hypothesis H_0 are independent and uniformly distributed, then the BH procedure guarantees control of the FDR at the nominal level α. Furthermore, Theorem 4 in (Storey et al., 2004) relaxes the independence assumption required by the BH procedure, providing theoretical guarantees for FDR control under the dependence of p-values. Since the aggregated p-values used in FTD approximately satisfy the required assumptions, applying the BH procedure ensures that the FDR is controlled at the desired level. Therefore, the FTD procedure controls the FDR at level α, as claimed. i.e.,

$$\text{FDR} := \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{D}_{\text{con}}|}{|\hat{\mathcal{S}}| \vee 1}\right] \leq \alpha.$$

□

Before giving the theorem 2, we introduce the following lemmas, which are useful for proving the power consistency.

Lemma 4 (Theorem 6 in (Storey et al., 2004), finite-sample version)**.** *Fix $t \in (0, 1]$ and let $q \in (0, 1)$. Then for any $\epsilon > 0$,*

$$\mathbb{P}\Big(\big|\widehat{\text{FDR}}(t) - \text{FDR}(t)\big| > \epsilon\Big) \leq 2\exp\Big(-2n \cdot \pi_0^2 t^2 \epsilon^2 \cdot \text{FDR}(t)^{-4}\Big).$$

This follows from applying Hoeffding's inequality to $\widehat{G}_n(t)$ and noting that $x \mapsto \pi_0 t/x$ is Lipschitz on bounded away-from-zero intervals.

Define the ideal threshold,

$$t^* := \sup\{t \in (0, 1] : \text{FDR}(t) \leq \alpha\},$$

and the empirical threshold (e.g. BH with plug-in π_0),

$$t_{BH} := \sup\{t \in (0, 1] : \widehat{\text{FDR}}(t) \leq \alpha\}.$$

We now state a finite-sample deviation bound for $|t_{BH} - t^*|$.

Lemma 5 (Finite-Sample Deviation of BH Threshold)**.** *Suppose $\text{FDR}(t)$ is strictly increasing and continuously differentiable in a neighborhood $[t^* - \epsilon, t^* + \epsilon]$, and its derivative satisfies $\text{FDR}'(t) \geq \alpha > 0$ in that region. Then for any $\epsilon > 0$ such that $t^* - \epsilon > 0$, we have,*

$$\mathbb{P}(t_{BH} < t^* - \epsilon) \leq 2\exp\left(-2n \cdot \frac{\pi_0^2 (t^* - \epsilon)^2 \epsilon^2}{\alpha^2}\right),$$

$$\mathbb{P}(t_{BH} > t^* + \epsilon) \leq 2\exp\left(-2n \cdot \frac{\pi_0^2 (t^* + \epsilon)^2 \epsilon^2}{\alpha^4}\right).$$

Proof. We analyze the lower tail; the upper bound follows similarly.

Suppose $t_{BH} < t^* - \epsilon$, i.e., there is no $t \in [t^* - \epsilon, t^*]$ such that $\widehat{\text{FDR}}(t) \leq \alpha$. Since $\text{FDR}(t) \leq \alpha$ on $[0, t^*]$, we must have,

$$\widehat{\text{FDR}}(t) > \alpha \geq \text{FDR}(t), \quad \forall t \in [t^* - \epsilon, t^*].$$

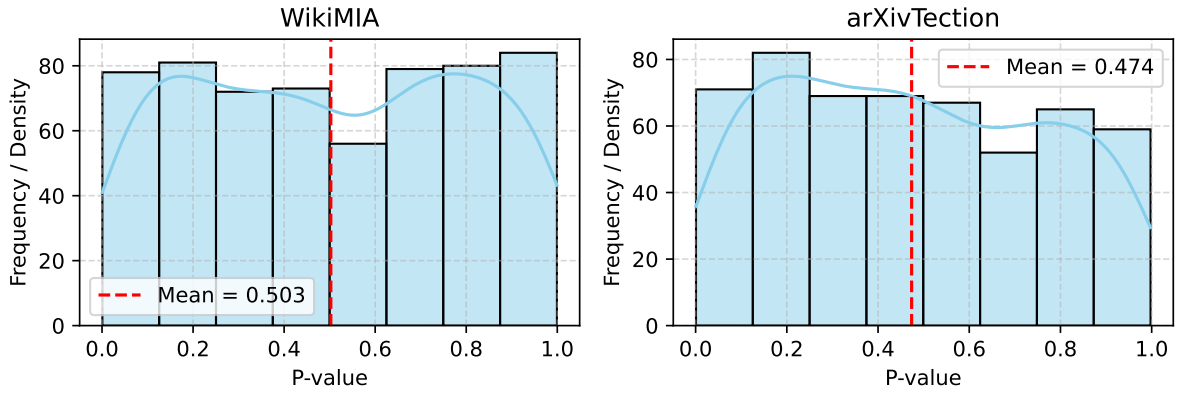

Figure 5: The p-value distribution of p_i^{agg} on datasets WikiMIA and arXivTection.

In particular, for $t = t^* - \epsilon$,

$$\widehat{\text{FDR}}(t) - \text{FDR}(t) > \alpha - \text{FDR}(t) \geq \alpha\epsilon,$$

where the last inequality comes from the first-order Taylor expansion,

$$\text{FDR}(t^*) - \text{FDR}(t^* - \epsilon) \geq \alpha\epsilon.$$

Therefore,

$$\mathbb{P}(t_{BH} < t^* - \epsilon) \leq \mathbb{P}\left(\widehat{\text{FDR}}(t^* - \epsilon) - \text{FDR}(t^* - \epsilon) > \alpha\epsilon\right),$$

which is bounded by Lemma 4,

$$\leq 2\exp\left(-2n \cdot \frac{\pi_0^2(t^* - \epsilon)^2(\alpha\epsilon)^2}{\text{FDR}(t^* - \epsilon)^4}\right)$$
$$\leq 2\exp\left(-2n \cdot \frac{\pi_0^2(t^* - \epsilon)^2\epsilon^2}{\alpha^2}\right)$$

□

Theorem 2 (Asymptotic Power Consistency of FTD)**.** *Assume that for each $k \in [K]$ and $Z_j \in \mathcal{D}_{\text{clean}}$, $\Pr(p_j^{(k)} \leq c) \geq 1 - \delta$ where $\delta = o(\sqrt{\log n/n})$ and c is a constant, and the density function of p_j^{agg} for $Z_j \in \mathcal{D}_{\text{clean}}$ has an upper bound $C_f > 0$, we have,*

$$\textit{Power} := \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{D}_{\textit{clean}}|}{|\mathcal{D}_{\textit{clean}}| \vee 1}\right] \geq 1 - C\sqrt{\log n/n}.$$

Proof. By Lemma 3, the aggregated p-value p_i^{agg}, computed from $p_i^{(1)}, \ldots, p_i^{(K)}$, is uniformly distributed on $[0, 1]$ under H_0. So $p_1^{agg}, \ldots, p_n^{agg}$ are drawn from the two-group mixture model,

$$p_i^{agg} \sim \pi_0 \cdot U[0, 1] + \pi_1 \cdot F_1, \quad \text{where } \pi_0 + \pi_1 = 1,$$

with F_1 being a continuous distribution supported on $[0, 1]$.

Define the empirical CDF,

$$\widehat{G}_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{p_i^{agg} \leq t\},$$
$$G(t) := \pi_0 t + \pi_1 F_1(t).$$

The true and estimated FDR curves are defined respectively as,

$$\text{FDR}(t) := \frac{\pi_0 t}{G(t)}, \quad \widehat{\text{FDR}}(t) := \frac{\pi_0 t}{\widehat{G}_n(t)}.$$

Define the ideal threshold,

$$t^* := \sup\{t \in (0, 1] : \text{FDR}(t) \leq \alpha\},$$

and the empirical threshold (e.g. BH with plug-in π_0),

$$t_{BH} := \sup\{t \in (0, 1] : \widehat{\text{FDR}}(t) \leq \alpha\}.$$

By the definition of power, we have,

$$\text{Power} := \frac{1}{|\mathcal{D}_{\text{clean}}|}\sum_{Z_i \in \mathcal{D}_{\text{clean}}} \mathbb{P}(p_i^{agg} \leq t_{BH})$$
$$= \Pr_{p_i^{agg} \sim F_1}(p_i^{agg} \leq t_{BH}).$$

Applying Lemma 5, we know that at probability $2\exp\left(-2n \cdot \frac{\pi_0^2(t^* - \epsilon)^2\epsilon^2}{\alpha^2}\right)$, $t_{BH} < t^* - \epsilon$ for any $\epsilon > 0$.

Then it has,

$$
\begin{aligned}
&\Pr_{p_i^{agg}\sim F_1}(p_i^{agg} \le t_{BH}) \\
&= 1 - \Pr_{p_i^{agg}\sim F_1}(p_i^{agg} > t_{BH}) \\
&= 1 - \Pr_{p_i^{agg}\sim F_1}(p_i^{agg} > t_{BH}, t_{BH} < t^* - \epsilon) \\
&\quad - \Pr_{p_i^{agg}\sim F_1}(p_i^{agg} > t_{BH}, t_{BH} \ge t^* - \epsilon) \\
&\ge 1 - \Pr_{p_i^{agg}\sim F_1}(p_i^{agg} > t^* - \epsilon, t_{BH} \ge t^* - \epsilon) \\
&\quad - \Pr_{p_i^{agg}\sim F_1}(p_i^{agg} > t_{BH}, t_{BH} < t^* - \epsilon) \\
&\ge \Pr_{p_i^{agg}\sim F_1}(p_i^{agg} \le t^* - \epsilon) - \Pr_{p_i^{agg}\sim F_1}(t_{BH} < t^* - \epsilon) \\
&\ge F_1(t^* - \varepsilon) - 2\exp\left(-2n \cdot \frac{\pi_0^2 (t^* - \epsilon)^2 \epsilon^2}{\alpha^2}\right).
\end{aligned}
$$

Moreover, under the assumptions in Theorem 2, by the definition of the weight where $\sum_{i=k}^{K} w_k = 1$, there exists $\ell \in [K]$ such that $\sum_{k=1}^{K} w_k \tan[(0.5 - p_i^{(k)})\pi] \ge \tan[(0.5 - p_i^{(\ell)})\pi]$. Then we have,

$$
\begin{aligned}
&F_1(t^*) = \Pr(p_i^{agg} \le t^*) \\
&= \Pr\left(\sum_{k=1}^{K} w_k \tan[(0.5 - p_i^{(k)})\pi] \ge \tan[(0.5 - t^*)\pi]\right) \\
&\ge \Pr\left(\tan[(0.5 - p_i^{(\ell)})\pi] \ge \tan[(0.5 - t^*)\pi]\right) \\
&\ge \Pr\left(\tan[(0.5 - p_i^{(\ell)})\pi] \ge \tan[(0.5 - t^*)\pi] \mid p_i^{(\ell)} < 0.5\right) \\
&\quad \cdot \Pr(p_i^{(\ell)} < 0.5)
\end{aligned}
$$

If $t^* \le 0.5$, as the function, $x \mapsto \tan\big[(0.5 - x)\pi\big]$ is strictly decreasing on $[0, 0.5)$, we have $p_i^{(\ell)} \le t^*$ is equivalent to,

$$
\tan\big[(0.5 - p_i^{(\ell)})\pi\big] \ge \tan\big[(0.5 - t^*)\pi\big].
$$

Therefore,

$$
\begin{aligned}
&\Pr\left(\tan[(0.5 - p_i^{(\ell)})\pi] \ge \tan[(0.5 - t^*)\pi] \mid p_i^{(\ell)} < 0.5\right) \\
&\quad \cdot \Pr(p_i^{(\ell)} < 0.5) \\
&= \Pr(p_i^{(\ell)} \le \min(t^*, 0.5)) \ge 1 - \delta.
\end{aligned}
$$

Otherwise, if $t^* > 0.5$, we have $\Pr\left(\tan[(0.5 - p_i^{(\ell)})\pi] \ge \tan[(0.5 - t^*)\pi] \mid p_i^{(\ell)} < 0.5\right) = 1$.

This leads to,

$$
\begin{aligned}
&\Pr\left(\tan[(0.5 - p_i^{(\ell)})\pi] \ge \tan[(0.5 - t^*)\pi] \mid p_i^{(\ell)} < 0.5\right) \\
&\quad \cdot \Pr(p_i^{(\ell)} < 0.5) \\
&\ge \Pr(p_i^{(\ell)} < c) \\
&\ge 1 - \delta.
\end{aligned}
$$

Combining together, we have,

$$
F_1(t^*) \ge 1 - \delta.
$$

Then note that $F_1(t^* - \varepsilon) \ge F_1(t^*) - f_1(t^*)\epsilon$. By the condition of $F_1(t^*) = 1 - \delta$ and $f_1(t) \le C$, we have $F_1(t^* - \varepsilon) \ge 1 - C\epsilon$. Putting together and taking $\epsilon = \frac{\pi_0 t^*}{\alpha}\sqrt{\log n/(2n)}$, we have,

$$
\begin{aligned}
\text{Power} &\ge 1 - \delta - C_f \epsilon \\
&\quad - 2\exp\left(-2n \cdot \frac{\pi_0^2 (t^* - \epsilon)^2 \epsilon^2}{\alpha^2}\right) \\
&= 1 - \delta - C_f \frac{\pi_0 t^*}{\alpha}\sqrt{\frac{\log n}{2n}} - \frac{2}{n},
\end{aligned}
$$

Since $\frac{2}{n} = o\left(\sqrt{\frac{\log n}{n}}\right)$ and all fixed factors can be absorbed into a universal constant, the bound simplifies to $1 - C\sqrt{\frac{\log n}{n}}$, where $C > 0$ is a constant depending on the testing level, the null proportion, and certain distributional characteristics. □

C Appendix C

C.1 Baselines

We provide detailed descriptions of the baseline methods used in our experiments:

- **PPL** (Li, 2023): This method uses the perplexity of a target model on a given input to infer whether the input was part of the training data. Lower perplexity often indicates higher likelihood of memorization.
- **Lowercase** (Carlini et al., 2021): This method calibrates the model's likelihood by comparing the perplexity of the original text with that of its lowercased version.
- **Zlib** (Carlini et al., 2021): This method uses compression entropy as a reference to calibrate the model's likelihood. The idea is that memorized or redundant content tends to be more compressible, and thus this ratio can help distinguish training data from non-training data.
- **Grad** (Hu et al., 2025): A gradient-based method that computes the norm of the gradient of the loss with respect to model parameters. Smaller gradient norms are often associated with training data.
- **Min-K%** (Shi et al., 2024): This method computes the average log-likelihood of the lowest $K\%$ tokens in a sequence. The K is set to 20.

Dataset	Type	Contaminated Data	Clean Data	Total
WikiMIA	Validation Set	258	237	495
	Test Set	603	552	1155
ArXivTection	Validation Set	228	236	464
	Test Set	534	550	1084
BBC Real Time	Validation Set	983	1003	1986
	Test Set	2293	2343	4636
MIMIR	Validation Set	1050	1050	2100
	Test Set	2450	2450	4900

Table 3: Statistics of the evaluation datasets used in our experiments.

- **Min-K%++** (Zhang et al., 2024b): An improved version of Min-K% that incorporates token-level calibration to enhance detection accuracy. The K is set to 20.

Moreover, to evaluate the effectiveness of our adaptive weighting strategy, we introduce three FTD variants: BH-Average, BH-Random, and BH-Max.

- **BH-Average**: discards the learned weights and instead averages the p-values uniformly.
- **BH-Max**: selects the best-performing detection method for each instance without fusion.
- **BH-Random**: assigns random weights to the detection methods.

C.2 Datasets

We evaluate our method on four benchmark datasets commonly used in training data detection: WikiMIA (Shi et al., 2024), ArXivTection (Duarte et al., 2024), BBC Real Time (Li et al., 2024), and MIMIR (Duan et al., 2024). These datasets contain both training data and non-training data, and are constructed to reflect realistic overlaps with pre-training corpora of large language models. Below, we briefly describe each dataset:

- **WikiMIA** (Shi et al., 2024): Contains Wikipedia event texts, where membership is determined based on publication timestamps. Events occurring before 2017 are treated as contaminated data, while those after 2023 are considered clean data.
- **ArXivTection** (Duarte et al., 2024): A benchmark dataset constructed from 50 research papers on arXiv, designed to evaluate pretraining data detection in scientific domains. Papers published before 2022 are labeled as contaminated data, while those from 2023 are labeled as clean data.
- **BBC Real Time** (Li et al., 2024): Comprises BBC news articles published between January 2017 and August 2024. Following the setup in (Shi et al., 2024), articles from 2017 are used as contaminated data, while those from 2024 serve as clean data.
- **MIMIR** (Duan et al., 2024): Constructed from the Pile dataset (Gao et al., 2020), where training samples are drawn from the train split and non-training samples from the test split. In our experiments, we select seven representative subsets—*DM Mathematics, GitHub, Pile CC, PubMed Central, ArXiv, HackerNews,* and *Wikipedia*—and report the averaged results.

For each dataset, we randomly select 30% as a validation set and use the remaining 70% for testing. The validation set is used to select decision thresholds or estimate the distribution of training data, while the test set is reserved for final evaluation. Detailed dataset statistics are provided in Table 3.

C.3 Implementation details

In real-world contamination detection, we require a threshold t to determine whether a data point is contaminated. To obtain this threshold, following the settings in (Shi et al., 2024; Hu et al., 2025; Zhang et al., 2024a), we randomly select 30% of the dataset as a validation set, while the remaining 70% is used as the test set. The optimal classification threshold is determined by maximizing detection accuracy on the validation set. In addition, our method requires access to a subset

Method	DM Mathematics			GitHub			Pile CC			PubMed Central		
	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑
PPL	0.380	0.489	0.594	0.316	0.303	0.581	0.235	0.549	0.690	0.319	0.414	0.610
Lowercase	0.425	0.317	0.541	0.270	0.317	0.600	0.234	0.654	0.727	0.323	0.497	0.630
Zlib	0.348	0.509	0.619	0.311	0.323	0.589	0.320	0.491	0.630	0.343	0.406	0.597
Grad	0.196	**0.963**	0.864	0.201	**0.820**	0.807	0.194	**0.880**	0.834	0.239	**0.711**	0.744
Min-K%	0.351	0.497	0.614	0.225	0.403	0.643	0.209	0.563	0.707	0.295	0.471	0.637
Min-K%++	0.161	0.880	0.856	0.142	0.809	0.837	0.188	0.777	0.799	0.238	0.703	0.741
FTD	**0.145**	**0.963**	**0.900**	**0.072**	0.740	**0.841**	**0.177**	0.863	**0.839**	**0.101**	0.611	**0.771**

Method	ArXiv			HackerNews			Wikipedia			Average		
	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑	FDR↓	Power↑	ACC↑
PPL	0.321	0.363	0.596	0.362	0.594	0.629	0.318	0.509	0.636	0.321	0.475	0.624
Lowercase	0.380	0.326	0.563	0.371	0.600	0.623	0.323	0.497	0.630	0.332	0.487	0.616
Zlib	0.390	0.300	0.554	0.371	0.509	0.604	0.316	0.537	0.644	0.343	0.439	0.606
Grad	0.197	0.806	0.804	0.292	**0.874**	0.757	0.255	0.877	0.789	0.225	0.847	0.800
Min-K%	0.269	0.466	0.647	0.302	0.700	0.699	0.272	0.520	0.663	0.274	0.550	0.660
Min-K%++	0.202	0.769	0.787	0.286	0.700	0.710	0.273	0.783	0.744	0.213	0.774	0.782
FTD	**0.146**	**0.812**	**0.836**	**0.158**	0.669	**0.771**	**0.117**	**0.879**	**0.880**	**0.131**	**0.802**	**0.834**

Table 4: Results on the challenging MIMIR benchmark with Pythia-2.8B

of known contaminated samples in order to estimate the p-values. Following a similar setting as in (Zhang et al., 2024a), we use the contaminated samples from the validation set for this purpose. For the baseline configurations, we follow the settings from their original papers. Specifically, the parameter K in both Min-K and Min-K++ is set to 20. Except for the *Effect of* α experiment, the value of α is fixed at 0.15 in all other settings. Furthermore, as our approach relies on integrating various training data detection methods, we incorporate all baseline methods for a comprehensive evaluation. All experiments are conducted on two NVIDIA A100 GPUs (40GB each) and a 16-core Intel Xeon Gold 6426Y CPU. All implementations are based on PyTorch.

C.4 Main results

The results for MIMIR are provided in the Table 4, which shows the same result of main paper. Specifically,

- *Finding 1 – FTD achieves the lowest FDR across all datasets and model settings, demonstrating its strong ability to suppress false positives.* Compared to the strongest baseline, FTD achieves the best FDR result, indicating a substantial improvement in precision when identifying clean data.
- *Finding 2 – FTD maintains competitive or superior statistical power in most settings.* It achieves near-perfect power on two datasets and performs robustly on the third, showing that it can effectively recall clean samples while maintaining low FDR.
- *Finding 3 – FTD consistently achieves the highest overall accuracy across different datasets and models.* This reflects its comprehensive effectiveness in both minimizing false detections and correctly identifying clean data, which is crucial for downstream applications.

Overall, our proposed method FTD consistently outperforms existing baselines across all evaluation metrics, demonstrating its robustness and effectiveness in mitigating the impact of contaminated data to support reliable model evaluation.

D Appendix D

D.1 LLM Evaluation Benchmark

We evaluate our method on four widely used large language model (LLM) evaluation benchmarks:

Original QA Pair	Formatted Prompt for Fine-tuning

SimpleQA
Q: What is the capital of France?
A: Paris.
	`### Question:`
	`What is the capital of France?`
	`### Answer:`
	`Paris.`

GPQA
Q: You come across an algorithm that gives the following output (written as input -> output):
AGG -> 115
TGCTGA -> 176
What value does ACAGTGACC give?
A: 315
	`### Question:`
	`You come across an algorithm that gives the following output (written as input -> output):`
	`AGG -> 115`
	`TGCTGA -> 176`
	`What value does ACAGTGACC give?`
	`### Answer:`
	`315`

TruthfulQA
Q: What tastes so good that cats ask for it by name?
A: Cats can't ask for anything by name.
	`### Question:`
	`What tastes so good that cats ask for it by name?`
	`### Answer:`
	`Cats can't ask for anything by name.`

ARC-C
Q: What do waves carry through objects?
(A) sound (B) light (C) energy (D) water
A: (C) energy
	`### Question:`
	`What do waves carry through objects?`
	`(A) sound (B) light (C) energy (D) water`
	`### Answer:`
	`(C) energy`

Table 5: Examples of question-answer pairs and their corresponding prompt format used for fine-tuning.

SimpleQA (Wei et al., 2024), GPQA (Rein et al., 2023), TruthfulQA (Lin et al., 2022), and ARC-C (Clark et al., 2018). These datasets are designed to measure complementary aspects of LLM. Below, we briefly describe each benchmark:

- **SimpleQA** (Wei et al., 2024): A benchmark designed to evaluate models on straightforward factual question answering. The dataset emphasizes clarity and unambiguous answers, making it suitable for measuring baseline factual knowledge.
- **GPQA** (Rein et al., 2023): A graduate-level benchmark focusing on expert-level knowledge across diverse scientific domains. Questions require precise reasoning and domain expertise, providing a challenging test for LLMs beyond basic factual recall.
- **TruthfulQA** (Lin et al., 2022): Consists of questions crafted to expose common misconceptions or false associations that language models may generate. The benchmark evaluates a models ability to provide factually correct responses while resisting the tendency to produce plausible but false statements.
- **ARC-C** (Clark et al., 2018): The Challenge subset of the AI2 Reasoning Challenge benchmark. It is composed of grade-school science exam questions requiring reasoning, inference, and integration of knowledge.

For each dataset, we follow the standard evaluation setup proposed in the corresponding papers.

D.2 Fine-tuning Experimental Setup

To investigate the practical applicability of FTD in mitigating the impact of test data contamination on model evaluation, we create a controlled synthetic contamination scenario. Specifically, we fine-tune two pre-trained models, **LLaMA-3.1-8B** and **Mistral-7B**, on four evaluation benchmarks: SimpleQA, GPQA, TruthfulQA, and ARC-C. For each dataset, we randomly sample **50%** of its test set and inject it into the fine-tuning data. The fine-tuned models are then re-evaluated on both the full benchmark test sets and the cleaned subsets produced by different contamination detection methods.

Data Preparation. To simulate evaluation data contamination, we randomly sample half (50%) of each benchmarks evaluation set and use it as supervised fine-tuning data. Each selected sample is a question-answer (QA) pair, which is formatted into a prompt suitable for instruction tuning. Examples of original QA pairs and their corresponding formatted prompts are shown in Table 5.

Training Configuration. The fine-tuning is conducted using the Hugging Face transformers and

LoRA Rank	LoRA α	Dropout	Epochs	Batch Size	LR	Scheduler
16	32	0.1	10	8	5e-5	Cosine

Table 6: Fine-tuning configuration.

Method	TruthfulQA				ARC-C			
	LLaMA-3.1-8B		Mistral-7B		LLaMA-3.1-8B		Mistral-7B	
	FDR↓	SRCC↑	FDR↓	SRCC↑	FDR↓	SRCC↑	FDR↓	SRCC↑
Contaminated	-	0.657	-	0.429	-	0.643	-	0.464
PPL	0.125	0.714	0.178	0.657	0.145	0.786	0.124	0.893
Lowercase	0.128	0.829	0.171	0.829	0.182	0.750	0.151	0.750
Zlib	0.120	0.829	0.165	0.657	0.125	0.929	0.101	0.964
Grad	0.128	0.657	0.198	0.771	0.109	0.964	0.075	**1.000**
Min-K%	0.133	0.715	0.208	0.829	0.134	0.857	0.113	0.893
Min-K%++	0.112	0.829	0.147	0.943	0.135	0.857	0.105	0.964
FTD	**0.074**	**1.000**	**0.087**	**1.000**	**0.061**	**1.000**	**0.070**	**1.000**

Table 7: Impact of different detection methods on LLM Evaluation for TruthfulQA and ARC-C. SRCC measures whether a detector can effectively remove contaminated data and thereby preserve ranking consistency in model evaluation; values closer to 1 indicate higher consistency.

peft libraries. Key hyperparameters are summarized in Table 6.

D.3 Impact of different detection methods on LLM Evaluation for TruthfulQA and ARC-C.

Based on Table 7, we draw the following observations: ❶ *Data contamination substantially distorts LLM evaluation results*. Under contaminated settings, the SRCC values drop markedly compared with the clean reference, indicating that contamination can severely alter model rankings and lead to unreliable or misleading evaluation outcomes. ❷ *Existing detection methods offer limited mitigation*. Methods such as PPL, Lowercase, Zlib, Grad, and Min-K variants partially reduce contamination effects, but they fail to consistently control FDR and often exhibit unstable or suboptimal SRCC across datasets and base models.❸ *FTD provides robust and reliable protection against contamination*. Across both TruthfulQA and ARC-C, and for both LLaMA-3.1-8B and Mistral-7B, FTD consistently achieves the lowest FDR while attaining perfect ranking consistency (SRCC = 1). These results underscore that strict FDR control is crucial for effectively mitigating contamination and ensuring fair, trustworthy LLM evaluation.

D.4 Results under Different Contamination Levels

In the original paper, we reported the performance of different training data detection methods under contamination. In this section, we further evaluate their robustness under increasing contamination levels (5%, 10%, and 25%) on the SimpleQA benchmark (Tables 8, 9, 10).

The model ranking on clean SimpleQA serves as the ground-truth reference (SimpleQA (Ground Truth), second column of the tables). To simulate contamination, we randomly select 5%, 10%, 25%, or 50% of the SimpleQA data to fine-tune LLaMA-3.1-8B, and then evaluate this contaminated model on the full SimpleQA benchmark (SimpleQA (Contaminated)). We subsequently apply different detection methods to filter contaminated samples from the evaluation set and re-rank the eight models. A detection method is considered more effective if the resulting ranking is closer to the ground truth.

For each setting, we report the FDR, defined as the proportion of contaminated samples incorrectly retained as clean among all samples predicted to be clean. From Tables 8–10, we draw the following findings: *Finding 1 — Data contamination substantially distorts model evaluation.* Contamination severely biases model rankings, and its impact increases with the contamination level. For example, LLaMA-3.1-8B shifts from rank 7 to rank 5 under 5% contamination, and further rises to rank 2 under 25% contamination. *Finding 2 — Existing detection methods provide partial but insufficient mitigation.* While all baseline methods reduce contamination effects compared to the fully contaminated setting, the contaminated model still ranks above its ground-truth position, indicating persistent residual contamination. *Finding 3 — FTD effectively restores fair model evaluation.* Across all contamination levels, FTD consistently achieves the lowest FDR and fully recovers the ground-truth ranking of LLaMA-3.1-8B, highlighting the critical role of strict FDR control.

Rank	SimpleQA (Ground Truth)	SimpleQA (contaminated)	PPL	Lowercase	Zlib	Grad	Min-K%	Min-K%++	FTD
1	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B
2	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini
3	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini
4	Gemini-1.5-Flash	Gemini-1.5-Flash	Gemini-1.5-Flash	Gemini-1.5-Flash	Gemini-1.5-Flash	Gemini-1.5-Flash	Gemini-1.5-Flash	Gemini-1.5-Flash	Gemini-1.5-Flash
5	Mistral-7B	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	Mistral-7B
6	Claude-3-Haiku	Mistral-7B	Mistral-7B	Mistral-7B	Mistral-7B	Mistral-7B	Mistral-7B	Mistral-7B	Claude-3-Haiku
7	**LLaMA-3.1-8B**	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	**LLaMA-3.1-8B**
8	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B
FDR	–	–	0.050	0.050	0.050	0.050	0.050	0.050	**0.016**

Table 8: Impact of different detection methods on LLM Evaluation. We simulate a data contamination scenario by fine-tuning LLaMA-3.1-8B on 5% of the SimpleQA training set. Various data detection methods are then employed to retain clean data for LLM evaluation. The effectiveness of a detection method is measured by how closely the resulting model leaderboard aligns with the ground truth ranking, the closer the match, the more effective the method.

Rank	SimpleQA (Ground Truth)	SimpleQA (contaminated)	PPL	Lowercase	Zlib	Grad	Min-K%	Min-K%++	FTD
1	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B
2	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini
3	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini
4	Gemini-1.5-Flash	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	Gemini-1.5-Flash	**LLaMA-3.1-8B**	Gemini-1.5-Flash	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	Gemini-1.5-Flash
5	Mistral-7B	Gemini-1.5-Flash	Gemini-1.5-Flash	**LLaMA-3.1-8B**	Gemini-1.5-Flash	**LLaMA-3.1-8B**	Gemini-1.5-Flash	Gemini-1.5-Flash	Mistral-7B
6	Claude-3-Haiku	Mistral-7B	Mistral-7B	Mistral-7B	Mistral-7B	Mistral-7B	Mistral-7B	Mistral-7B	Claude-3-Haiku
7	**LLaMA-3.1-8B**	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	**LLaMA-3.1-8B**
8	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B
FDR	–	–	0.099	0.084	0.097	0.083	0.095	0.087	**0.032**

Table 9: Impact of different detection methods on LLM Evaluation. We simulate a data contamination scenario by fine-tuning LLaMA-3.1-8B on 10% of the SimpleQA training set. Various data detection methods are then employed to retain clean data for LLM evaluation. The effectiveness of a detection method is measured by how closely the resulting model leaderboard aligns with the ground truth ranking, the closer the match, the more effective the method.

Rank	SimpleQA (Ground Truth)	SimpleQA (contaminated)	PPL	Lowercase	Zlib	Grad	Min-K%	Min-K%++	FTD
1	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B	LLaMA-3-70B
2	GPT-4o-mini	**LLaMA-3.1-8B**	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini
3	o1-mini	GPT-4o-mini	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini	o1-mini
4	Gemini-1.5-Flash	o1-mini	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	**LLaMA-3.1-8B**	Gemini-1.5-Flash
5	Mistral-7B	Gemini-1.5-Flash	Gemini-1.5-Flash	Gemini-1.5-Flash	Gemini-1.5-Flash	Gemini-1.5-Flash	Gemini-1.5-Flash	Gemini-1.5-Flash	Mistral-7B
6	Claude-3-Haiku	Mistral-7B	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku	Claude-3-Haiku
7	**LLaMA-3.1-8B**	Claude-3-Haiku	Mistral-7B	Mistral-7B	Mistral-7B	Mistral-7B	Mistral-7B	Mistral-7B	**LLaMA-3.1-8B**
8	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B	LLaMA-3.2-3B
FDR	–	–	0.162	0.152	0.166	0.114	0.171	0.133	**0.046**

Table 10: Impact of different detection methods on LLM Evaluation. We simulate a data contamination scenario by fine-tuning LLaMA-3.1-8B on 25% of the SimpleQA training set. Various data detection methods are then employed to retain clean data for LLM evaluation. The effectiveness of a detection method is measured by how closely the resulting model leaderboard aligns with the ground truth ranking, the closer the match, the more effective the method.